Researching an Enhanced Multimodal Learning Framework for Improved Inter-Modal and Intra-Modal Alignment

Zijun Long

PhD thesis

Supervised by Dr. Richard Mccreadie & Dr. Gerardo Aragon Camarasa

School of Computer Science
College of Science and Engineering
University of Glasgow

Nov 2024

Abstract

In recent decades, machine learning research has predominantly focused on single-modal data. However, the emergence of multimodal data, such as images or videos accompanied by text, particularly on social media platforms, has underscored the importance of advancing multimodal learning. This thesis centers on multimodal learning, exploring ways to enhance the performance of multimodal models—specifically those utilizing vision and language modalities. It aims to improve the understanding and integration of multimodal data, thereby boosting performance in downstream tasks such as crisis response, robotics, cross-modal retrieval, and recommendation.

In this thesis, we argue that enhancing shallow inter-modal and intra-modal alignment in existing multimodal approaches can improve performance across different tasks by enabling deeper alignment. To address this, we introduce a novel multimodal learning framework, named MCA, designed to improve multimodal learning performance while maintaining flexibility across various downstream tasks. The framework comprises three core components: Mixture-of-Modality-Experts (MoME), Contrastive Learning Techniques, and Adapter Methods, each offering unique functionalities.

Firstly, the Mixture-of-Modality-Experts (MoME) component is designed to manage a diverse range of input modalities and improve inter-modal alignment. Recent years have seen a significant shift towards multimodal learning, yet many existing models are mere amalgamations of single-modal models, using fusion layers to merge separate vision and language models. This method often leads to shallow alignments and can compromise the effectiveness of multimodal models. To overcome these limitations, MoME enables a unified model architecture, incorporating a modality-specific expert system adept at processing multimodal data (notably vision and language) for a variety of downstream tasks, such as classification and image-text retrieval. Benefiting from this design, MoME has the ability to process different combinations of input, such as unimodal, multimodal, or mixed.

Secondly, to enhance intra-modal and inter-modal alignment and bolster performance across both unimodal and multimodal contexts, we researched several innovative contrastive learning techniques. Initially, our research focused on label-aware contrastive learning for image models, resulting in a robust encoder for image inputs. Subsequently, we introduced an Optimized Learning Fusion strategy, termed CLCE, designed to refine the optimization process by integrating the cross-entropy loss function with the contrastive learning loss function. Furthermore, we devel-

oped a debiased contrastive learning approach aimed at mitigating label noise within the contrastive learning framework, thereby further enhancing model performance. Collectively, these methodologies fortify the contrastive learning component of our multimodal learning framework, significantly deepening inter-model alignment and augmenting overall effectiveness.

Thirdly, to address the challenges of efficiency and practicality associated with large-scale models, we have developed an innovative approach to transfer learning utilizing adapters. As the size of Multimodal Large Language Models (MLLMs) increases, their adaptation to specific tasks becomes more complex, primarily due to heightened computational and memory requirements. Traditional fine-tuning methods, while effective, are resource-intensive and necessitate extensive, task-specific training. Although various adaptation methods have been proposed to mitigate these issues, they often result in inadequate inter-modal alignment, compromising the models' overall effectiveness. In response to these challenges, we present the MultiWay-Adapter (MWA), a novel method equipped with an 'Alignment Enhancer'. This feature significantly improves inter-modal alignment, facilitating efficient model transferability with minimal tuning. Consequently, the MWA emerges as a highly efficient and effective method for adapting MLLMs, substantially enhancing their utility across a broader range of applications.

Each proposed approach within the framework is rigorously assessed using one or more specially curated datasets for that component. This evaluation includes a detailed analysis of the approaches, identifying suitable settings for their deployment, and providing insights into their performance characteristics.

This thesis has made contributions to the field of multimodal learning by enhancing both intra-modal and inter-modal alignment, improving computational efficiency, and validating the proposed MCA framework in real-world applications. Our evaluations provide multiple pieces of evidence for improved alignment and enhanced performance across various metrics in evaluated datasets, supporting our thesis statement. These advancements pave the way for future research and development in creating more effective and efficient multimodal systems.

Furthermore, this thesis extends to the comprehensive evaluation and optimization of the proposed framework across various domains, such as crisis response, robotics, and cross-modal retrieval. Insightful findings are drawn from an extensive series of experiments that cover the proposed framework of multimodal learning. The results presented within this thesis highlight the improvements our framework contributes to both the overarching benchmarks of multimodal learning and a wide array of downstream applications.

We applied multimodal learning to crisis response, addressing the limitation of prior works that primarily use single-modality content. This thesis examines the importance of integrating multiple modalities for crisis content categorization. We design a multimodal learning framework that fuses textual and visual inputs, leveraging both to classify content based on specific tasks. Using the CrisisMMD dataset, we demonstrate effective automatic labeling with an average of 88.31% F1 performance across relevance and humanitarian category classification tasks.

We also analyze the success and failure cases of unimodal and multimodal models.

The second application of our multimodal learning framework is in robotic vision, which requires tasks like object detection, segmentation, and identification. Integrating specialized models into a unified vision pipeline poses engineering challenges and costs. Multimodal Large Language Models (MLLMs) have emerged as effective backbones for various tasks. Leveraging the pre-training capabilities of MLLMs simplifies the framework, reducing the need for task-specific encoders. The large-scale pre-trained knowledge in MLLMs allows for easier fine-tuning and superior performance in robotic vision tasks. We introduce the RoboLLM framework, equipped with a BEiT-3 backbone, to handle all visual perception tasks in the ARMBench challenge. RoboLLM outperforms existing baselines and significantly reduces the engineering burden of model selection and tuning.

The third application in this thesis is text-to-image retrieval, which finds relevant images based on text queries. This is crucial for digital libraries, e-commerce, and multimedia databases. While multimodal models show state-of-the-art performance in some retrieval tasks, they struggle with large-scale, diverse, and ambiguous real-world needs due to computational costs and injective embeddings. To address this, we present the two-stage Coarse-to-Fine Index-shared Retrieval (CFIR) framework for efficient large-scale long-text to image retrieval. The first stage, Entity-based Ranking (ER), handles query ambiguity using a multiple-queries-to-multiple-targets paradigm. The second stage, Summary-based Re-ranking (SR), refines rankings with summarized queries. We also propose a specialized Decoupling-BEiT-3 encoder for both stages, enhancing computational efficiency with vector-based similarity inference. Evaluations on the AToMiC dataset show that CFIR outperforms existing MLLMs by up to 11.06% in Recall@1000, while reducing training and retrieval times by 68.75% and 99.79%, respectively.

Acknowledgements

This thesis was made possible by the tremendous support I received from a wide variety of people during the course of my PhD.

First and foremost, I would like to thank my supervisor, Richard McCreadie, and my co-supervisor, Gerardo Aragon Camarasa. Their guidance taught me how to integrate and build upon prior works from various fields to tackle new tasks and design sound experiments. Without their meticulous attention to detail, this thesis would not have been possible.

I am also deeply grateful to my family, who continuously encouraged me and provided financial support, enabling me to pursue my studies at the University of Glasgow. I am especially grateful to my girlfriend, Wanxing Li, for her unwavering support and meticulous care in every possible way.

I would like to extend my heartfelt thanks to my outstanding collaborators, Paul Henderson, Lipeng Zhuang, George Killick, Xuri Ge and Muhammad Imran. I have benefited immensely from their insightful perspectives and wisdom. I cherish the time we have spent together and have greatly enjoyed our discussions. Completing so many papers would not have been possible without their joint efforts.

Additionally, I would like to thank the TerrierTeam research group, who have provided invaluable support over the past four years (and counting). Special thanks go to Iadh Ounis, Craig Macdonald, Jingmin Huang, Yaxiong Wu, Xiao Wang, Graham McDonald, Ting Su, Zeyan Liang, Xi Zhang, Sean MacAvaney, Zaiqiao Meng, Zeyuan Meng, Jake Laver, Jack McKechnie, Lubingzhi Guo, Sarawoot Kongyoung, Alexander Pugantsov, Aleksandr Petrov, Andrew Parry, Gan Wang, Jinyuan Fang, Javier Sanz-Cruzado Puig.

Lastly, I want to express my gratitude to the many friends I met in Glasgow, especially Xicheng Li. Their support and the fun times we shared gave me the strength to endure the challenging periods of my PhD.

Contents

Chapter 1

Introduction

1.1 Introduction

Historically, machine learning has predominantly been unimodal, focusing either on text [159, 182, 286] or images [21, 42, 150]. Unimodal learning, which relies on a single type of data, faces several limitations compared to multimodal learning [13, 17, 117]. It often provides limited context and information, reducing accuracy and robustness, particularly when the single modality lacks sufficient details or contains noise. Unimodal models struggle with generalization across different tasks and domains and offer a limited perspective, which is inadequate for understanding complex phenomena [117]. They are also less effective in handling real-world data, which is inherently multimodal, and cannot exploit interactions between different modalities, crucial for nuanced tasks [13]. These constraints make unimodal learning less effective and practical for diverse and complex applications.

On the contrary, multimodal learning, which integrates multiple types of data such as text and images, has demonstrated considerable enhancements in a plethora of machine learning tasks, as substantiated by numerous studies [184] [14]. This approach uncovers valuable insights that may be hidden within images rather than solely in textual data. By concentrating exclusively on one modality, potential nuances and insights could remain unexplored.

Moreover, with recent advances in deep learning, particularly in the fields of computer vision [81] and language modeling [46], there has been a heralding of a new era of potential for multimodal learning [63, 104, 127, 225, 267]. The application of sophisticated visual recognition models, trained on datasets like ImageNet, enables the extraction of insights from images to be combined with textual data. These advancements in unimodal deep neural models enable the fusion of embeddings from different modalities, facilitating the creation of powerful multimodal systems. This fusion is invaluable in applications such as crisis response, where messages may comprise solely text, images, or a combination of both.

A critical challenge in this field is to achieve a comprehensive understanding of multimodal content, necessitating the simultaneous analysis of both textual and visual data for holistic interpretation. The foundation for effective interpretation in multimodal learning lies in generating high-quality embeddings for each modality. Unlike unimodal learning, multimodal learning requires producing embeddings for all modalities within a unified semantic space. This process involves not only intra-modal alignment (within a single modality) but also inter-modal alignment (between different modalities). For example, in the domain of crisis response, a tweet might express a request for assistance through text, accompanied by an image depicting a scene of devastation. Here, while the text delineates the nature of the request, the accompanying image provides essential locational context. Although there have been ventures into this direction, most previous works involve the use of separated dual backbones: separate encoders for text and images. This approach has seen some success in multimodal contexts but is fraught with limitations. The integration of modalities in these designs is often *superficial (shallow inter-modal alignment)*, limited to a few densely connected layers at best [13]. Moreover, The dual-encoder architecture inherently increases the model's size, leading to longer training and processing times [231]. These models are also predominantly trained on datasets pairing text with images, limiting their flexibility for single-modal downstream tasks. This situation is exacerbated by the comparatively smaller size of text-image paired datasets relative to their single-modal counterparts, reducing the available training data and potentially compromising model performance.

Entering 2022, large language models emerged as powerful encoders for text input, excelling in a wide range of tasks [3, 50, 123, 252, 275]. This advancement also benefited multimodal learning research by providing more powerful encoders for different modalities, leading to the development of Multimodal Large Language Models (MLLMs) [123, 252]. While MLLMs offer superior performance compared to other neural models, they also present notable shortcomings, particularly their size [143, 145, 146, 148, 149, 271]. As MLLMs grow, adapting them to specialized tasks becomes increasingly challenging due to high computational and memory demands. Traditional fine-tuning methods are costly, requiring extensive task-specific training. This issue underscores the need for efficient learning methods for MLLMs. Although some efficient adaptation methods aim to reduce these costs, they often suffer from shallow inter-modal alignment, which significantly reduces model effectiveness [37, 222, 228, 279].

1.2 Thesis Statement

In this thesis, we argue that enhancing shallow inter-modal and intra-modal alignment in existing multimodal approaches can improve performance across different tasks by enabling deeper alignment. To address this, we hypothesize that utilizing the three pivotal components proposed in this thesis—Mixture-of-Modality-Experts (MoME), Contrastive Learning Techniques, and Adapter Methods—will enhance both inter-modal and intra-modal alignment, leading to significantly improved effectiveness and efficiency of multimodal models across a range of tasks. The core of our framework comprises the MoME component, which leverages shared transformer block parameters to enhance computational efficiency and facilitate deeper modality integration, contrastive learning methods to improve fusion and alignment for better generalization across data types, and adapter-based transfer learning techniques to address the practical challenges of using large models efficiently. Our framework is expected to outperform state-of-the-art models, such as LXMERT and VisualBert, in vision-language benchmarks including Visual Question Answering (VQA) and Natural Language for Visual Reasoning (NLVR). Beyond standard benchmarks, our multimodal learning framework will undergo extensive testing in real-world applications, focusing on three key downstream tasks: crisis response, image-text retrieval, and robotics. We anticipate that our enhanced multimodal learning framework will consistently solve real-world problems and exhibit superior performance in effectiveness and efficiency across these diverse domains.

1.3 Contributions

This thesis contributes to the field of multimodal learning in two key ways. Firstly, we demonstrate that shallow inter-modal and intra-modal alignment compromises the quality of the produced embeddings, thereby limiting overall performance. To tackle this, we introduce an innovative multimodal learning framework, MCA, designed to address the issue of shallow alignment, including both inter-modal alignment and intra-modal alignment, thereby enhancing effectiveness and efficiency, in Chapter 4. Secondly, we research and optimize our proposed framework across three distinct domains, demonstrating its superior performance and broad applicability, from Chapter 7 to 9.

1.3.1 Contributions of the Proposed MCA Framework

This framework is structured around three principal components: contrastive learning techniques, Mixture-of-Modality-Experts (MoME), and adapter methods. Each component is designed to offer distinct functionalities that address specific challenges identified in the multimodal learning domain. These challenges include deepening inter-modal and intra-modal alignment, processing a diverse array of input modalities and ensuring the framework's efficiency and practicality for real-world applications. Importantly, while each component of our framework targets specific obstacles outlined in Section 1.1, they are synergistically integrated, ensuring robust effectiveness and efficiency.

Thereafter, we undertake an exhaustive examination of each component, wherein we propose, develop, and rigorously assess innovative methodologies aimed at enhancing their functionality. Below, we outline our contributions for each component, demonstrating how they collectively contribute to advancing the state of multimodal learning.

Firstly and most importantly, as stated in our thesis statement, we address the issue of shallow inter-modal and intra-modal alignment and improve performance across both unimodal and multimodal contexts by researching several innovative contrastive learning techniques. Given that contrastive learning techniques were originally designed for visual learning and that visual encoders play a crucial role in multimodal learning for providing robust intra-modal alignment, we began our research by enhancing the effectiveness of image models through contrastive learning methods. Initially, we focused on label-aware contrastive learning for image models during the fine-tuning stage, resulting in a robust encoder for image inputs. Subsequently, we introduced an optimized learning fusion strategy, termed CLCE, which refines the optimization process by integrating the cross-entropy loss function with the contrastive learning loss function. Furthermore, recognizing that label noise is common in large training datasets, particularly for multimodal learning datasets, we developed a debiased contrastive learning approach aimed at mitigating label noise within our contrastive learning framework, thereby further enhancing model performance. Collectively, these methodologies fortify the contrastive learning component of our multimodal learning framework, significantly deepening inter-modal alignment and augmenting overall effectiveness.

Secondly, the MoME component is designed to manage a diverse range of input modalities and improve inter-modal alignment. Recent years have seen a significant shift towards multimodal learning, yet many existing models are mere amalgamations of single-modal models, using several fusion layers to merge separate vision and language models. This method often leads to shallow alignments and can compromise the effectiveness of multimodal models. To overcome these limitations, MoME enables a unified model architecture, incorporating a modality-specific expert system adept at processing multimodal data (notably vision and language) for a variety of downstream tasks, such as classification, recommendation, and image-text retrieval. Moreover, previous studies have tended to concentrate on either unimodal or multimodal inputs, neglecting the real-world scenario where inputs can vary—ranging from solely text to combinations of text and images. The MoME approach not only accommodates such variability but also leverages larger unimodal datasets for pretraining, thereby overcoming the constraints posed by the smaller size of multimodal datasets. This broader training foundation enriches the model with more extensive pretrained knowledge, leading to improved performance. Within the MoME framework, specialized experts are deployed for different modal inputs—for instance, a text expert for textual data and a vision-language expert for combined visual and textual information.

Thirdly, to address the challenges of efficiency and practicality associated with large-scale models, especially for Multimodal Large Language Models (MLLMs), we have developed an innovative approach to transfer learning utilizing adapters. Indeed, as the size of MLLMs increases, their adaptation to specific tasks becomes more complex, primarily due to heightened computational and memory requirements. Traditional fine-tuning methods, while effective, are resource-intensive and necessitate extensive, task-specific training. Although various adaptation methods have been proposed to mitigate these issues, they often result in inadequate inter-modal alignment, compromising the models' overall effectiveness. In response to these challenges, we present the MultiWay-Adapter (MWA), a novel framework equipped with an 'Alignment Enhancer'. This feature improves both inter-modal and intra-modal alignment, facilitating efficient model transferability with little tuning. Consequently, the MWA emerges as a highly efficient and effective method for adapting MLLMs, substantially enhancing their utility across a broader range of applications.

1.3.2 Contributions from Practical Applications

In addition to developing the proposed framework, another major contribution of this work lies in its application across three distinct domains: crisis response, robotics, cross-modal retrieval. We meticulously tailored the proposed framework for optimized performance within each domain, conducting comprehensive evaluations. These efforts resulted in achieving superior performance benchmarks compared to the existing state-of-the-art solutions.

Specifically, we study the application of multimodal learning in the domain of crisis response. We first begin our research on utilizing text data and aim at build robust text encoder to subsequent research on multimodal learning. Social media platforms, like Twitter, are increasingly used by billions of people internationally to share information. As such, these platforms contain vast volumes of real-time multimedia content about the world, which could be invaluable for a range of tasks such as incident tracking, damage estimation during disasters, insurance risk estimation, and more. By mining this real-time data, there are substantial economic benefits, as well as opportunities to save lives. The COVID-19 pandemic attacked societies at an unprecedented speed and scale, forming an important use-case for social media analysis. However, the amount of information during such crisis events is vast and information normally exists in unstructured and multiple formats, making manual analysis very time consuming. Most prior works in this area use machine learning to categorize single-modality content (e.g., text or images), with few studies jointly utilizing multiple modalities. In chapter 7, we examine the importance of integrating multiple modalities for crisis content categorization and how inter-modal alignment affect the performance. Specifically, we design a framework for multimodal learning that fuses textual and visual inputs, leverages both, and classifies the content based on the specified task. Through evaluation using the CrisisMMD dataset, we demonstrate that automatic labeling with multimodal data is effective with deep inter-modal alignment, achieving an average F1 performance of 88.31% across two important tasks (relevance and humanitarian category classification), while also analyzing the success and failure cases of unimodal and multimodal models.

Beyond the crisis response, we optimized our proposed multimodal learning framework in robotic domain. Robotic vision applications often necessitate a wide range of visual perception tasks, such as object detection, segmentation, and identification. While there have been substantial advances in these individual tasks, integrating specialized models into a unified vision pipeline presents significant engineering challenges and costs. With the integration of MLLMs, our proposed framework can handle various robotic vision perception tasks. We argue that leveraging contrastive learning methods and the pre-training capabilities of MLLMs enables the creation of a simplified framework, thus mitigating the need for task-specific encoders. Specifically, the large-scale pretrained knowledge in MLLMs allows for easier fine-tuning to downstream robotic vision tasks and yields superior performance. We introduce the RoboLLM framework, equipped with a BEiT-3 backbone, to address all visual perception tasks in the ARMBench challenge—a large-scale robotic manipulation dataset about real-world warehouse scenarios. RoboLLM not only outperforms existing baselines but also substantially reduces the engineering burden associated with model selection and tuning. This achievement demonstrates that the enhanced performance of the multimodal learning framework can also contribute to performance improvements in unimodal tasks. For example, RoboLLM solves the object identification task with a 97.8% recall@1.

Another important application for multimodal learning is cross-modal retrieval, particularly text-to-image retrieval. This task, which involves finding relevant images based on a text query, is crucial in various use-cases such as digital libraries, e-commerce, and multimedia databases. Although multimodal models demonstrate state-of-the-art performance in some retrieval tasks, they face limitations in handling large-scale, diverse, and ambiguous real-world retrieval needs due to computational costs and the injective embeddings they produce. Therefore, this thesis presents a two-stage Coarse-to-Fine Index-shared Retrieval (CFIR) framework based on our proposed multimodal learning framework, designed for fast and effective large-scale long-text to image retrieval. The first stage, Entity-based Ranking (ER), addresses long-text query ambiguity by employing a multiple-queries-to-multiple-targets paradigm, facilitating candidate filtering for the next stage. The second stage, Summary-based Re-ranking (SR), refines these rankings using summarized queries. Additionally, we propose a specialized Decoupling-BEiT-3 encoder, optimized for handling ambiguous user needs in both stages, enhancing computational efficiency through vector-based similarity inference. Evaluation on the AToMiC dataset reveals that CFIR surpasses existing MLLMs by up to 11.06% in Recall@1000, while reducing training and retrieval times by 68.75% and 99.79%, respectively.

In summary, the central contributions of this thesis include addressing the issue of shallow inter-modal and intra-modal alignment in multimodal learning, as outlined in the thesis statement, by introducing an effective and efficient multimodal learning framework. This framework is applied to four impactful and distinct domains: crisis response, robotics, cross-modal retrieval. This research draws from a diverse range of experiments across various domains, including computer vision, natural language processing, and neural models, to validate and refine the framework. We meticulously tailored the framework for optimized performance in each domain and conducted comprehensive evaluations. The experimental results presented in this thesis demonstrate the framework's effectiveness and efficiency, highlighting its wide applicability in enhancing performance across different tasks.

1.4 Origins of the Material

Portions of the research presented in this thesis have been previously disseminated through conference proceedings and journal articles. The publications listed below are directly relevant to the themes of this thesis and underpin the research elaborated upon in various chapters:

1. **Zijun Long**, Lipeng Zhuang, George Killick, Richard McCreadie, Gerardo Aragon Camarasa, Paul Henderson. Understanding and Mitigating Human-Labelling Errors in Supervised Contrastive Learning, *The 18th European Conference on Computer Vision*, Full Paper, 2024. (Core A*, h5-index 238) [link]

2. **Zijun Long**, Xuri Ge, Richard Mccreadie and Joemon Jose. CFIR: Fast and Effective Document-To-Image Retrieval for Large Corpora, 2023. *International ACM SIGIR Conference on Research and Development in Information Retrieval*, Full Paper, 2024. (Core A*, h5-index 103) [link]

3. **Zijun Long**, George Killick, Richard McCreadie, Gerardo Aragon Camarasa. RoboLLM: Robotic Vision Tasks Grounded on Multimodal Large Language Models, *IEEE International Conference on Robotics and Automation*, Full Ppaer, 2024. (Core A*, h5-index 119) [link]

4. **Zijun Long**, George Killick, Richard McCreadie, Gerardo Aragon Camarasa. MultiWay-Adapater: Adapting large-scale multimodal models for scalable image-text retrieval. *IEEE International Conference on Acoustics, Speech, and Signal Processing*, main, 2024. (h5-index 123) [link]

5. **Zijun Long**, Richard McCreadie, Gerardo Aragon Camarasa, Zaiqiao Meng. LaCViT: A Label-aware Contrastive Training Framework for Vision Transformers. *IEEE International Conference on Acoustics, Speech, and Signal Processing*, main, 2024. (h5-index 123) [link]

6. **Zijun Long**, George Killick, Richard McCreadie, Gerardo Aragon Camarasa, Zaiqiao Meng. CLCE: An Approach to Refining Cross-Entropy and Contrastive Learning for Optimized Learning Fusion, *The 27th European Conference on Artificial Intelligence*, Full Paper, 2024. (h5-index 36) [link]

7. **Zijun Long**, Richard McCreadie, Muhammad Imran. CrisisViT: A Robust Vision Transformer for Crisis Image Classification. *20th International Conference on Information Systems for Crisis Response and Management*, main, 2023. (h5-index 21)[link]

8. **Zijun Long**, Richard McCreadie. University of Glasgow Terrie Team at the TREC 2023 AToMic Track. *TREC* 2023.

9. **Zijun Long**, Richard McCreadie. Is Multi-Modal Data Key for Crisis Content Categorization on Social Media? *19th International Conference on Information Systems for Crisis Response and Management*, main, 2022. (h5-index 21)[link]

10. **Zijun Long**, Richard McCreadie. Automated Crisis Content Categorization for COVID-19 Tweet Streams. *18th International Conference on Information Systems for Crisis Response and Management*, main, 2021. (h5-index 21) [link]

11. **Zijun Long**, Xiaohang Wang, Yingtao Jiang, Guofeng Cui, Li Zhang, Terrence Mak. Improving the Efficiency of Thermal Covert Channels in Multi-/many-core Systems. *Design, Automation and Test in Europe*, main, 2017. (h5-index 49) [link]

- Chapter 4: The architectures of proposed framework and Mixture-of-modality-expert design were published in [2,3,4].

- Chapter 5: Three techniques we proposed to improve the effectiveness of contrastive learning have been published in [1,5,6].

- Chapter 6: The efficient transfer learning method of adapter has been published in [4].

- Chapter 7: The application in crisis response of our proposed framework have been published in [7,9,10].

- Chapter 8: The application in robotic vision of our proposed framework have been published in [3].

- Chapter 9: The application in cross-modal retrieval of our proposed framework have been published in [2].

1.5 Thesis Outline

The structure of this thesis is outlined as follows:

Chapter 2 lays the foundational knowledge required for understanding multimodal learning. This chapter traces the evolution from traditional machine learning techniques to advanced deep learning methodologies in both unimodal and multimodal contexts, complemented by a review of relevant literature. It also introduces the concepts of contrastive learning and parameter-efficient learning approaches as they pertain to Multimodal Large Language Models (MLLMs) and Large Language Models (LLMs).

Chapter 3 presents the benchmarks utilized to assess the performance of multimodal models, alongside detailed descriptions of the datasets employed in the pretraining of our multimodal models.

Chapter 4 describes the architecture of our proposed multimodal learning framework, including the Mixture-of-modality-expert model and our enhancements to it.

Chapter 5 delves into the advancements in contrastive learning methods, focusing on their role in developing a robust image encoder for multimodal learning.

Chapter 6 outlines our contributions towards designing parameter-efficient techniques, specifically adapters, for the fine-tuning of MLLMs and LLMs.

Chapter 7 examines the application of our proposed multimodal framework in crisis response.

Chapter 8 investigates the implementation of our proposed multimodal framework in robotics.

Chapter 9 discusses the utilization of our proposed multimodal framework in cross-modal retrieval.

Chapter 10 provides concluding remarks and summarizes the contributions of this thesis. Additionally, it discusses the limitations and presents several future directions that build upon the foundation laid in this thesis.

Chapter 2

Background and Related Work

As discussed in Section 1.2, this thesis focuses on addressing the issue of shallow intra-modal alignment in multimodal learning to improve performance. Specifically, we focus on multimodal learning using textual and image data. Text and images are among the most widely used and rich sources of information in various applications. Textual data provides detailed, context-rich information that can be used to understand complex narratives, sentiments, and specific details. Images, on the other hand, offer visual context and can capture details that text alone might miss, such as spatial relationships and visual attributes. By combining these two modalities, we aim to leverage their complementary strengths to enhance the performance of multimodal learning systems. This combination is particularly useful in applications such as crisis response, where both visual and textual information are crucial for accurate and timely decision-making. A schematic view of the standard multimodal learning framework, along with its components and relevant research areas, is presented in Figure 2.1. In a typical multimodal learning framework, separate encoders are used to process language and visual inputs independently. The embeddings from each modality are then fused using fusion methods, which generally employ contrastive learning to ensure intra-modal alignment and place the embeddings in the same semantic space. Furthermore, we not only evaluate the proposed framework, MCA, on general benchmarks but also optimize and test it in real-world scenarios.

We begin with a description of traditional machine learning methods in language learning in Section 2.1, followed by a discussion of more recent progress in deep learning methods for language learning in Section 2.2. Next, we cover another important modality in this thesis, visual data, starting with traditional machine learning methods in Section 2.3, and then proceeding to deep learning approaches in Section 2.4. After reviewing the works in unimodal learning, we provide a comprehensive background on multimodal learning in Section 2.5. We further introduce the background of parameter-efficient learning approaches in Section 2.7. Subsequently, we offer a detailed investigation of works from different domains: crisis response in Section 2.9, cross-modal retrieval in Section 2.11, and robotic vision in Section 2.10.

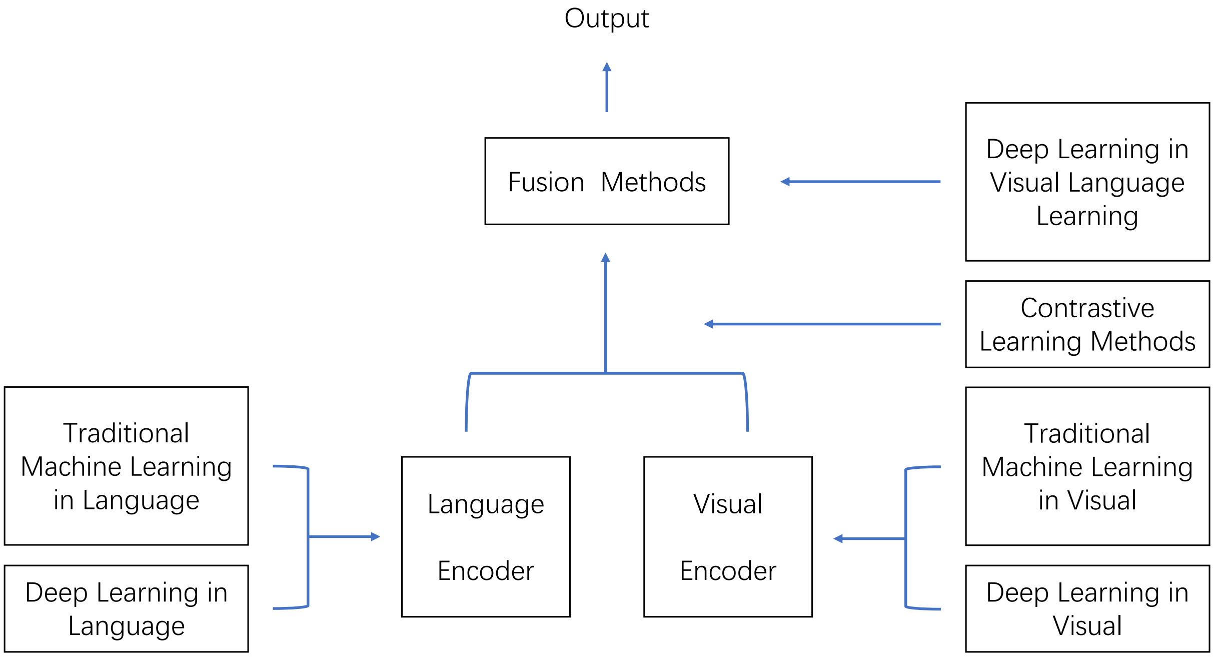

Output
Fusion Methods
Deep Learning in Visual Language Learning
Contrastive Learning Methods
Traditional Machine Learning in Language
Deep Learning in Language
Language Encoder
Visual Encoder
Traditional Machine Learning in Visual
Deep Learning in Visual

Figure 2.1: The general structure of a multimodal learning framework.

2.1 Traditional Machine Learning in Language

Machine learning models have been instrumental in the domain of text mining for an extended period. Roughly before 2012, traditional Machine Learning techniques had become essential tools for processing language data, which had been classified as Natural Language Processing. Traditional methodologies in this realm frequently employ bag-of-words embeddings [286], a technique where text is represented by the presence or absence of words from a predetermined dictionary. Despite their widespread use, these models are inherently limited by their inability to reconcile semantic discrepancies between the dictionary terms and the actual text [75, 243]. Furthermore, they treat words in isolation, which significantly constrains the depth of linguistic meaning that can be captured, as contextual nuances and word relationships within the text are largely ignored [75, 243].

As we transitioned around 2013, the field began to move away from bag-of-words models in favor of shallow word-embedding techniques. This paradigm shift was largely facilitated by the advent of pre-trained neural network models that transform individual words into vectors, which are then aggregated to form a comprehensive vector representation of the text [159]. Unlike their predecessors, these models are adept at capturing semantic relationships between words, thereby addressing the semantic mismatch issue inherent in bag-of-words approaches. By designing these models to assign similar vector representations to words with analogous meanings, they significantly enhance the model's ability to understand and process natural language in a more nuanced and context-aware manner.

Expanding upon this, it's essential to acknowledge the progression to more sophisticated language models that build on the foundation laid by shallow embeddings. The development of models such as Word2Vec [159] and GloVe (Global Vectors for Word Representation) [182] marked significant milestones in natural language processing. These models not only improved the efficiency and quality of embeddings but also introduced the ability to capture a broader range of semantic and syntactic relationships between words.

In summary, the evolution from traditional bag-of-words models to contemporary neural learning-based approaches in language processing reflects significant advancements in our ability to model and understand natural language. These developments underscore a shift towards deep learning models, which are more contextually aware, semantically rich, and syntactically sophisticated models, enabling a broad spectrum of applications in natural language processing and beyond.

2.2 Deep Learning in Language

The advent of deep learning ushered in a transformative era in natural language processing (NLP), marked by the development of more sophisticated embedding models capable of capturing the nuanced meanings embedded within sequences of text. Early works by Simonyan and Zisserman [214], Razavian et al.[209], and Antonellis et al.[69] laid the groundwork for these advancements, significantly increasing the complexity of models to enhance their semantic understanding capabilities.

This evolution paved the way for the emergence of modern pre-trained deep neural language models. Such models, characterized by their extensive complexity and generality, are designed for embedding sequences of text. They are pre-trained on vast corpora, enabling them to perform a multitude of natural language processing tasks with unprecedented efficiency and accuracy. Today, the landscape of text categorization is dominated by these pre-trained neural language models, with GPT (Generative Pre-trained Transformer) [191] and BERT (Bidirectional Encoder Representations from Transformers) [46], along with their derivatives such as DistilBERT [205] and RoBERTa (A Robustly Optimized BERT Pretraining Approach) [139], leading the charge.

Deep neural Language models before BERT [46], such as OpenAI's GPT [191] and ELMo [183], utilize unidirectional architectures for pre-training, limiting their ability to fully understand context, especially for token-level tasks like question answering which require understanding from both directions of a sentence. GPT [191], for example, can only process previous tokens with its left-to-right design, while ELMo [183] adopts task-specific architectures for a feature-based approach. These constraints led to the development of BERT [46], which introduces a bidirectional approach through a masked language modeling (MLM) and next sentence prediction for enhanced text-pair representations. BERT's architecture allows for significant improvements in efficiency, robustness, and performance across a broad range of natural language processing tasks, achieving state-of-the-art results, including a notable increase in the GLUE benchmark score. Its unified framework facilitates easy application to various tasks with minimal adjustments and short fine-tuning periods with the help of transfer learning methods. Moreover, BERT demonstrates that scaling model size can lead to substantial performance gains, even with limited data, setting new precedents in NLP task performance. Despite its remarkable capabilities, BERT has limitations, such as a maximum input length of 512 tokens and the requirement of substantial computational resources, which can challenge its deployment on standard GPU memory.

Around 2021, large models for languages based on transformer architecture, such as BERT and RoBERTa, became commonly known as large language models (LLMs). These models showcased unprecedented capabilities in language generation, comprehension, and various NLP tasks, setting a new standard for LLMs. Exemplified by OpenAI's GPT (Generative Pre-trained Transformer) series, which further pushed the boundaries by demonstrating the power of unsupervised pre-training on vast text corpora, followed by fine-tuning on specific tasks. GPT-3, with 175 billion parameters, showcased unprecedented capabilities in language generation, comprehension, and various NLP tasks, setting a new standard for LLMs. Building on the success of GPT-3, OpenAI introduced ChatGPT [173], a conversational model based on the GPT-3 [19] architecture, fine-tuned for dialogue generation. ChatGPT quickly became popular for its ability to engage in coherent and contextually relevant conversations, making it suitable for applications like customer service, virtual assistance, and interactive entertainment.

The evolution continued with the introduction of GPT-4 [174], which brought even greater advancements in model size, performance, and versatility. GPT-4 demonstrated enhanced reasoning abilities, improved context understanding, and more accurate generation of text across a broader range of topics. These advancements have significantly enhanced the integration of LLMs in various applications, from sophisticated chatbots and virtual assistants to complex content generation and multimodal learning frameworks, continuously driving the field forward with ongoing research and development.

The construction of a robust text encoder is pivotal in developing a robust multimodal framework, serving as a fundamental for enhancing the intra-modal alignment. The advancements in models like GPT and BERT have not only revolutionized text categorization but have also set new standards for the development of models capable of understanding and interpreting the intricate interplay between text and other modalities, such as images and videos. Hence, our research commences with the foundational task of building an effective text encoder, leveraging the advanced capabilities of these deep learning models to enhance our understanding and processing of natural language.

2.3 Traditional machine learning in Visual

Before the advent of deep learning, machine learning in image recognition was a field rich with diverse methodologies and approaches aimed at understanding and classifying visual content. This section provides a comprehensive overview of the key techniques and milestones in image recognition before deep learning became the dominant methods. Through examining these foundational concepts and methods, we gain insight into the evolution of image recognition and the groundwork it laid for the deep learning revolution.

Image recognition, a cornerstone task in the field of computer vision, involves identifying objects, people, text, and other elements within digital images. Prior to deep learning's prevalence, researchers developed various algorithms and models to tackle this challenge, relying on hand-crafted features and classical machine learning techniques. We explore these pre-deep learning methodologies, shedding light on their significance and how they paved the way for today's advanced neural network-based solutions.

The cornerstone of traditional image recognition was the extraction of meaningful features from images. These features, which describe various properties such as edges, textures, and shapes, were crucial for the subsequent classification tasks. There are several milestone studies:

- Edge Detection: Techniques such as the Canny, Sobel, and Prewitt operators were fundamental in identifying edges in images, serving as primary features for object detection and recognition tasks [21].

- Scale-Invariant Feature Transform (SIFT): Lowe et al. introduced SIFT algorithm which was a breakthrough, allowing for the detection and description of local features in images that were invariant to scale and rotation, significantly aiding in object recognition and matching tasks [150].

- Histogram of Oriented Gradients (HOG): The HOG descriptor, introduced by Dalal et al. [42], was pivotal for object detection, particularly in pedestrian detection, by capturing edge and gradient information distributed across localized regions of an image.

Despite the successes of these methods, there were significant challenges, including the labor-intensive process of designing and selecting features, the difficulty in handling high variability within object classes, and the limited capacity to process complex scenes with multiple interacting objects.

The era preceding deep learning in image recognition was marked by significant innovation and development. Techniques such as SIFT, HOG, and SVMs were instrumental in advancing the field, setting the stage for the deep learning revolution. While deep learning has since overshadowed these methods, understanding the evolution of image recognition provides valuable insights into the complexities of visual perception and the ongoing quest to mimic human-level understanding in machines.

2.4 Deep learning over Images

Similar to text-based categorization discussed above, supervised image categorization is also dominated by deep learned models. In this case, the item embedding step takes the pixel data from the image as input, which is fed into a deep neural network that extracts some meaning from the image. As before, the deep neural network will have been pre-trained on a number of tasks, such as image classification or object detection [66, 198, 201], enabling it to learn what is important to extract and embed from the image. A separate model or classification layer can then be trained on-top of the deep neural model. On the other hand, unlike for text, there are currently two competing architectures: convolutional neural networks (CNNs); and transformers. CNNs have traditionally been the dominant neural network type used for image classification. The reason for this is the input dimensionality is much higher for images than text (there are more pixels in an image than words in a sentence), meaning effective dimensionality reduction is key. Architecturally CNNs are advantaged here, as their convolutional structure forces them to find the parts of the image that matter and discard the rest, enabling them to better generalize to unseen examples. As a result, pre-trained CNNs are popular choices as baselines, such as ResNet152 [81] and VGG [214].

Transformers were first proposed by [242] and have remained popular for NLP tasks. Several state-of-the-art models such as BERT [46] and GPT [192] come pre-trained on large datasets, often spanning several related tasks. These models can then be 'tuned' with new examples to enable them to be applied for other tasks, while still leveraging much of the pre-trained network. One of the key components of these models is the attention mechanism, which makes the embedding of each token dependent on every other token (and those tokens' relative positions) in the input sequence, enabling these models to dynamically adapt the embedding of a token based on the context it appears within. However, in the computer vision domain, a naive application of the attention mechanism dramatically increases computational complexity, because the number of pixels in an image is much higher than the typical number of words in a sentence. Therefore, early works experimented with different methods to minimise the cost of attention, while maintaining its benefits. For instance, [180] applied self-attention to only the local neighborhood for each query pixel, while [256] applied attention to only small parts of the image instead. The recent ViT model [49], introduced in 2020, was the first vision transformer model that was able to apply attention globally with minimal modifications to the transformer architecture, but was still incredibly expensive to train (over 2,500 TPUv3-core days). In 2021, a new MAE model was proposed that addresses the speed problem through the use of a masked autoencoder architecture with an encoder-decoder pre-training schema, which enables MAE to learn how to reconstruct the original image based on only partial observations of that image [115]. This setting markedly reduces the number of pixels that need to be fed into the transformer resulting in faster training, and is the best current solution to this challenge. Explorations in various pretraining methods, including SimMIM [262] and Data2vec [11], further advanced the field.

Building a robust image encoder is critical for providing better embeddings for visual input and addressing the issue of shallow intra-modal alignment in multimodal learning. Given the dominant performance of transformer-based image models, we began our research on enhancing these models to develop an improved multimodal learning framework. However, despite these advancements, a common challenge remains: the limited transferability of transformer-based image models, particularly when using cross-entropy loss for fine-tuning [49, 292]. This motivated us to explore solutions in Chapter 5, focusing on improving model transferability through contrastive learning methods.

2.5 Deep learning in Visual-Language Learning

As noted earlier, most prior works in machine learning are unimodal, considering only a single modality (text or images). However, several relevant multimodal tasks have been investigated, including visual reasoning [227] and visual question answering [70]. In this thesis, we focus on vision-language learning, which involves both visual and language inputs.

Most vision-language models for multimodal tasks train text and image encoders separately, feeding the outputs into a final classification model [104, 267]. We believe this late-stage modality combination results in shallow intra-modal alignment, as stated in the thesis statement. Beyond the shallow intra-modal alignment issue, other problems such as efficiency and speed arise, since simply extracting input features requires significantly more computation than the multimodal interaction steps. A less explored alternative is to use a unified architecture, where a single embedding model takes both modalities as input, such as VISUALBERT [127] or VL-BERT [225]. In these models, the modality combination and interaction occur early and throughout the entire model. Recent work has indicated that this early interaction between modalities can be beneficial to performance [283]. Furthermore, [231] proposed a model named LXMERT, which applies a cross-attention mechanism to different modalities, enabling a new way to fuse information from various modalities and achieving state-of-the-art results.

In this thesis, we aim to investigate how these architectures affect the performance of vision-language learning and analyze their pros and cons. This analysis provides insights into how to develop an enhanced multimodal learning framework that improves intra-modal alignment.

2.6 Contrastive Learning

Contrastive learning has emerged as a cornerstone in representation learning, particularly in self-supervised and multimodal contexts. This section reviews its development, focusing on key advancements, applications, and challenges.

The development of contrastive learning traces back to early explorations by Becker [16]. This approach aims to differentiate similar items from dissimilar ones within an embedding space. Specifically, contrastive learning is a learning paradigm that compares groups of $\langle$item, prediction$\rangle$ pairs, rather than considering each pair in isolation. The fundamental idea is to communicate to the model the degree and direction of error in its representation by contrasting the embeddings of correctly and incorrectly predicted examples. This process teaches the model to make the embeddings of examples belonging to a single class more similar, while simultaneously pushing apart the embeddings of examples from different classes.

Contrastive learning has demonstrated remarkable efficacy in improving deep learning model performance across various domains [142, 143], including sentence [67, 134] and audio representation learning [294], with its most notable impact observed in image recognition tasks, exemplified by SimCLR [29] and other studies [144, 145, 149].

While integrating label information into contrastive learning has been explored, as in [109], these efforts have primarily remained confined to the pre-training phase and have not been extended to vision transformers. Despite parallel advancements in both fields, the integration of label-aware contrastive learning within the fine-tuning stage of vision transformers remains unexplored. Therefore, in Section 5.2, we address this gap by pioneering the application of contrastive learning during the fine-tuning phase of vision transformers, thereby enhancing their transferability. This enhancement in vision encoders eventually improves the performance of multimodal learning when integrating them into the framework. Additionally, these contrastive learning techniques can be used in training multimodal learning frameworks to enhance intra-modal alignment, such as the alignment of semantic information from different modalities.

2.6.1 Contrastive Learning in Multimodal Contexts

In multimodal learning, contrastive learning plays a pivotal role in aligning representations across modalities. Models such as CLIP [190] and ALIGN [101] have successfully employed contrastive objectives to create shared embedding spaces for text and images. These methods align modality-specific representations by treating paired data (e.g., an image and its caption) as positive pairs and unrelated data as negative pairs. This alignment enables impressive zero-shot and few-shot capabilities in cross-modal retrieval and classification tasks.

2.6.2 Negative Mining in Contrastive Learning

The exploration of negative samples, particularly hard negatives, in contrastive learning has emerged as a critical yet relatively underexplored area. While the significance of positive sample identification is well-established, recent studies have begun to unravel the intricate role of hard negatives. The potential of hard negative mining in latent spaces has been validated in numerous studies [39, 120, 226, 258, 265, 287]. These studies highlight the pivotal role of hard negatives in enhancing the discriminative capability of embeddings. In the contrastive learning domain, [40] tackled the challenge of discerning true negatives from a vast pool of candidates by approximating the true negative distribution. Later, [203] applied hard negative mining to unsupervised contrastive learning, resulting in a framework where only a single positive pair is utilized in each iteration of the loss calculation. However, these approaches still present limitations, such as inaccurately identifying positive and negative samples and only using one positive pair, which harms the performance of contrastive learning. [103] expand upon the concept of hard negative mining within a supervised framework. Their approach employs a consistent threshold-based dot product for identifying "hard" samples. However, determining an appropriate threshold remains challenging, as it varies significantly across different datasets and even within individual mini-batches. Moreover, their methodology does not tackle the dependency on large batch sizes, which is a critical limitation on performance and applicability.

By utilizing negative mining techniques, there is significant potential to improve the effectiveness of contrastive learning methods, which in turn enhances the performance of multimodal learning. These techniques can be employed in training multimodal learning frameworks to further improve intra-modal alignment.Therefore, in Section 5.3, we propose a new objective name CLCE, which builds on these foundational insights, aiming to synergize the strengths of contrastive learning with Cross-Entropy (CE), particularly by employing hard negative mining guided by label information. CLCE employs a dynamic and adaptive strategy to assign weights to "hard" samples in each minibatch, offering a more refined approach compared to previous studies. Additionally, CLCE achieves superior performance to CE without relying on large batch sizes.

2.6.3 Human Labelling Errors and Contrastive Learning

Human-labelling errors are prevalent in many datasets used for supervised learning, especially for large datasets where eliminating errors is impractical [169, 276]. Mislabeled examples can lead to overfitting in models, with larger models being more susceptible [53, 77, 169, 219, 280]. Robust learning from noisy labels is thus crucial for improving generalization. Methods include estimating noise transition matrices [138, 269, 272], regularization [89, 99, 282], and sample re-weighting [200, 248].

Contrastive learning, particularly in unsupervised visual representation learning, has evolved significantly [16, 24, 29, 33, 73, 278]. However, the absence of label information can result in positive samples in negative pairs, potentially leading to detrimental effects on the representations learned. SCL leverages labeled data to construct positive and negative pairs based on semantic concepts of interest (e.g. object categories). It ensures that semantically related points are attracted to each other in the embedding space. Khosla et al. [109] introduce a SCL objective inspired by InfoNCE[172], which can be considered a supervised extension of previous contrastive objectives—e.g. triplet loss [206] and N-pairs loss [137]. Despite its efficacy, the impact of label noise and the importance of hard sample mining in SCL are often overlooked.

We build on their work in Section 5.4 by devising a strategy that not only mitigates the impact of human-labelling errors but also enhances the performance of the SCL objective.

Previous noise-mitigation works in SCL primarily target synthetic labelling errors, often excluding human-mislabeled samples through techniques like specialized selection pipelines [178] and bilevel optimization [99]. Recently, Sel-CL [128] introduced a non-linear projection head for intra-sample similarity analysis to identify confident samples, and TCL [94] employs a Gaussian mixture model with entropy-regularized cross-supervision. Despite these advances, such methods risk overfitting on these confident pairs, particularly in the presence of human-labelling errors, reducing their effectiveness in real-world scenarios as shown in our analysis (Sec. 5.4.4). They also focus on artificially noisy datasets with synthetic noise rates up to 80%, an unrealistic scenario in practice. Our proposed method in Section 5.4 instead specifically targets human-labelling errors in common image datasets, optimizing SCL performance without excluding all of these errors. In summary, mitigating human labeling errors further enhances the effectiveness of contrastive learning methods, which in turn improves the performance of multimodal learning through better intra-modal alignment.

In summary, contrastive learning has evolved into a robust framework for self-supervised and multimodal learning. Its application in creating aligned, generalizable representations across diverse data types continues to expand, making it a critical area of ongoing research in machine learning.

2.7 Parameter Efficient Learning Approaches

Recently, increasing model size has been shown to be an effective strategy for improving performance. Models such as BEiT-3 [252] and BLIP-2 [123], with up to 1.9 billion and 12.1 billion parameters, respectively, have set new state-of-the-art results in multimodal tasks such as Visual Question Answering. However, their application to specialized downstream tasks is often limited by computational constraints [143, 145, 146, 148, 149, 271]. For instance, the requirement for large GPU memory in full fine-tuning limits their adaptations for specialized tasks on commodity hardware, e.g., 45 GB for full fine-tuning of the BEiT-3 Large model.

The challenge of computational efficiency in fine-tuning MLLMs has given rise to Parameter-Efficient Transfer Learning (PETL) methods. These are broadly categorized into partial parameter updates [222] and modular additions [181, 228]. The former is resource-intensive and model-specific, while the latter adds new modules to architectures, updating only these components. However, most studies only focus on unimodal tasks in domains such as vision [197], text [86] or audio [108, 233, 241], neglecting multimodal tasks. A few works [37, 222, 228, 279] target multimodal tasks but suffer from shallow inter-modal alignment.

Therefore, in this thesis, to enhance the applicability of our proposed framework and enable evaluation in real-world scenarios, we research and propose a parameter-efficient learning method named MultiWay-Adapter in Chapter 6. This method is designed for efficient MLLMs transfer learning and improved inter-modal alignment.

2.8 Target Use-cases

In this thesis, we target four critical domains to demonstrate the effectiveness of our proposed multimodal learning framework: crisis response, cross-modal retrieval, robotic vision, and recommendation. Crisis response involves analyzing real-time multimedia content to provide timely and accurate information during emergencies, enhancing decision-making and resource allocation. Cross-modal retrieval focuses on retrieving relevant images based on textual queries, a key task in digital libraries, e-commerce, and multimedia databases. Robotic vision addresses the need for advanced visual perception in robotics, encompassing tasks such as object detection, segmentation, and identification in complex environments. Lastly, recommendation systems are essential for personalizing content in online platforms, leveraging diverse data sources to improve user experiences. These domains are explored in detail in subsequent sections, highlighting the related work and our contributions in each area.

2.9 Crisis Response

Social media is increasingly seen as a critical platform for emergency response, as a channel to gather and analyze urgent information during a crisis [113, 118, 192]. For example, information in Twitter has previously been shown to be useful for detecting infectious disease [270]. Indeed, within social media, a common task for emergency responders is to filter and categorize information on these platforms with the aim of finding actionable content for the response effort [116, 165, 186]. However, due to the large volumes of information published on social media platforms, a tremendous amount of time and effort would be needed to perform this task manually, which is simply impractical. To solve this, supervised machine learning solutions have been proposed to do this task automatically [156, 212]. For example, Twitter data can help many emergency departments [165] and public health agencies [62] to predict disease spread. Moreover, geographically tagged social media content has shown to be a valuable tool for tracing and mapping disease outbreaks [257]. Meanwhile, the TREC Incident Streams (TREC-IS) track examined the automatic categorization of social media posts into 25 information types [156]. These works have shown that supervised machine learned approaches for identifying actionable information from social media are feasible, and to some degree, effective, although more work is still needed in this area [63]. Therefore, we consider crisis response as a importance application and we propose our solutions to this in Chapter 7. Therefore, we regard crisis response as a critical and meaningful application and present our proposed solutions to fill the gap in Chapter 7.

In later sections we provide a brief overview of recent papers in crisis informatics, as well as past works within TREC Incident Streams track that are relevant to our investigation. Next, we provide a brief summary of relevant works from the literature on content categorization for pandemics, the TREC-IS initiative, and machine learning over social media data.

Social Media During Emergencies

Social media is a new but critical platform for relevant party to gather and analyse urgent information, especially like Twitter. Information extraction from social media platforms like Twitter is a recent but increasingly critical problem. Information collected via Twitter has previously been shown to be useful for detecting infectious disease both spatially and temporally [270], HIV/AIDS [62], seasonal influenza [165] and Ebola [110]. Indeed, within social media streams, a common task for emergency responders is to classify documents based on the information they contain. Twitter data, as a popular data source, can help many emergency departments [165] and public health agencies [62] to predict disease spread. Moreover, geographically tagged social media content has shown to be a valuable tool for tracing and mapping disease outbreaks [257]. However, up until now, few agencies actively take advantage of these resources.

TREC-IS Pilot Effort in 2020

The Text Retrieval Conference (TREC) Incident Streams track (denoted TREC-IS) is a public data challenge that aims to tackle current issues with automatically extracting actionable content from social media during crises. TREC-IS provides an excellent opportunity to develop and evaluate AI systems for crisis response. At a high level, participant TREC-IS systems can perform two tasks: classifying tweets by information type, and ranking tweets by criticality. For both tasks, given an event, a participating system receives a stream of filtered, event-relevant tweets and an ontology of information types from TREC-IS. The goal of that system is to produce tweet-level labels and priority ratings, which they then submit for evaluation. TREC-IS has run editions in 2018, 2019 and 2020. Importantly for this work, in response to the global COVID-19 pandemic the 2020 editions of TREC-IS introduced a COVID-19 sub-task and provided labelled tweets for evaluation. In particular, TREC-IS 2020 defines information ‘types’ to represent categories of information that emergency response officers might find interesting, for TREC-IS 2020 COVID-19 task (Task 3), the information types are as follows:

1. GoodsServices: The user is asking for a particular service or physical good.
2. InformationWanted: The user is requesting information.
3. Volunteer: The user asks people to volunteer to help the response effort.
4. EmergingThreats: The user reports problems may cause loss of life or damage.
5. NewSubEvent: The user reports a new occurrence that officers need to respond to.
6. ServiceAvailable: The user says that he or someone else is providing a service.

7. Advice: The author provides some advice to the public.

To capture the importance a given message has to emergency response officers, TREC-IS defines four information criticality labels: low, medium, high, and critical, where high- and critical-level messages require prompt or immediate review and potentially action by an emergency manager. For instance, examples of critical information might include calls for search and rescue, emergence of new threats (e.g., a infected patient), or calls for emergency medical care.

Machine Learning Approaches

For reference, we consider classical approaches to be those that rely on either bag-of-words or shallow embeddings to represent tweet text. Indeed, according to a 2019 review conducted by McCreadie et al. for TREC-IS Task 1 and 2 (Crises), classical classifiers can still be very competitive and robust, even when comparing to state-of-the-art deep neural network models [155], although that study did not cover pandemic-type events like COVID-19.

In contrast, recently, pre-trained deep neural language models have become popular as they are very effective methods to encode meaning contained within sequences of text [6, 209, 214]. These models replace the traditional bag-of-words or shallow word embeddings used by classical models. At the time of writing, the most widely used neural language model is the transformer BERT model [46] and its subsequent variants. For the purposes of classification, BERT and similar models can be tuned to produce a numeric vector representing a text sequence, which can then be passed to a traditional classification model. While models like BERT are widely seen as superior to more traditional text representation approaches [261], they are not yet commonly used in production systems due to their high computational cost and the need for dedicated GPU acceleration.

It is worth noting that models like BERT can be re-trained or tuned to make them more effective for particular domains or tasks. In the COVID-19 space, [164] recently produced an updated BERT model by re-training it over a COVID-19 twitter dataset. However, given the small gains in down-stream performance reported (around 0.03 F1) and the large cost of retraining the model, it is unclear whether the benefits are worth the effort and cost.

Tackling Class Imbalance

A concern with content classification for COVID-19 is the class imbalance [97, 111, 255]. For TREC-IS Task 3, there are 7 categories of interest, where only a small proportion of the tweets belong to each class. This is a challenge when training models, as there are few positive examples to learn from, leading to model bias towards the majority class. Moreover, from a task perspective, emergency responders are more sensitive to failures regarding the positive class, as this represents potentially useful information not being surfaced to the user.

A common approach for solving class imbalance is to balance the number of positive or negative samples used for training. For example, by down-sampling the majority class, over-sampling the minority class, or using a combination of the two [25, 51, 111, 153]. Alternatively, a number of learning methods that intrinsically account for class imbalance have been proposed, e.g. [111]. However, these require larger numbers of positive samples to be effective than are available for this task, hence we employ sampling methods here. Deep neural network models also suffer from imbalanced training data [92, 98]. Hard sample mining is a technique that has been exploited in computer vision to solve the class imbalance, e.g. for tasks such as object detection [58, 213], image categorisation [171], and unsupervised representation learning [253]. Hard sample mining focuses on selecting samples that represent difficult to classify, as they carry more discriminative power for the classifier to learn from.

TREC-IS Participant Systems

Participants to TREC-IS 2020 have developed a range of initial solutions for the COVID-19 task, where details can be found in the associated technical reports (known as ‘notebooks’) provided by TREC.[1] For instance, [246] experimented with two multi-task transfer learning approaches, one is an encoder-based model like BERT, while the other one leverages a sequence-to-sequence transformer, such as T5. These models do not explicitly attempt to counteract the problems of class imbalance in the crisis data. In contrast [211] tackle this problem via the automatic generation of additional examples via a synonym-augmentation strategy using the CrisisMMD dataset as a ground truth. Notably, this work applies a VGG model to classify images attached to the tweets, enabling both text and image data to be considered, which to some degree alleviates the issues with class imbalance.

2.9.1 Image content from Social Media platforms for Crisis Response

Social media is increasingly seen as a critical information and communication platform during emergencies, as a channel to gather and analyze urgent information during a crisis [118, 192, 236]. However, the majority of prior work in this space has focused on analysing textual content posted to these platforms rather than imagery [96]. On the other hand, recent works have begun to explore the value-add that crisis images posted to social media platforms can bring, as well as how to minimise the costs associated to image analysis through AI automation. For example, [167] demonstrated that crisis images on social media can be used for a variety of humanitarian aid activities (such as identifying areas in need of goods and services). Meanwhile, [1] showed that social media images are helpful for damage assessment during flooding events, while [43] illustrated that geotagged images can be used to identify affected regions in need of aid. Functionally, crisis image analysis can be seen as a classification or tagging problem, where a human

[1] `https://trec.nist.gov/proceedings/proceedings.html`

or machine needs to analyse the image and then assign a label or labels to that image.

To-date the crisis image domain has largely focused on four image classification use-cases:

- *Disaster Type Detection*: The high-level classification of the type of disaster depicted within an image, such as an earthquake, fire or flood.
- *Informativeness/Usefulness*: The classification of images to determine whether it contains some form of valuable information for an emergency responder. Typically represented as a binary informative/not informative classification.
- *Humanitarian Categories*: This form of classification is focused on what is happening within the image, where the goal is to identify images that are relevant to different types of humanitarian response activities. Common humanitarian categories include images of affected individuals, images of infrastructure or utility damage, or images of people needing rescue.
- *Damage Severity*: Finally, one common use for crisis images is to judge the severity of damage in a particular area, which is useful for response prioritization or damage costing purposes. The damage severity task mainly targets three levels: severe damage, mild damage, and little or no damage.

2.9.2 Multimodal Learning in Crisis Response

In the crisis informatics domain, there has been recent interest in multimodal learning for tasks such as tagging of social media data [2]. For instance, [170] proposed a multimodal deep learning pipeline to aid disaster response using a late interaction between a custom CNN for text and VGG16 [214] for images. They use joint representation for text and image, which means two parallel and separated architectures for text modality and image modality. For image modality, they use famous VGG16 to extract high-level features in images. For text, they define a Convolution Neural Network (CNN) with five hidden layers and different filters. But they only try their model on a small subset of Crisismmd dataset. They choose this subset because it is same label of those samples for text and image. In another word, they avoid the problem of handling different label or prediction for text and image. Furthermore, how to represent features of text and image is a critical research question for multimodal learning. However, this paper only simply concatenate output from two separated models for text and image, respectively. And they put concatenated vectors into another hidden layer and then use SoftMax as output function. This obviously is not the best solution. Meanwhile, [295] proposed a similar late interaction model, combining FastText [105] for text and VGG16 [214] for images. Thus, in this thesis, we examine the impact of our proposed multimodal framework in Chapter 4, and explore how to optimize the framework to achieve better performance in Section 7.2.

2.10 Robotic Vision

Robotic vision is a crucial downstream task chosen for this thesis due to its significant impact on various industries, including manufacturing, logistics, and service robotics. The ability of robots to accurately perceive and interpret their environment is fundamental to performing complex tasks such as object detection, segmentation, and manipulation. Advancements in robotic vision can lead to more efficient and autonomous robotic systems, reducing the need for human intervention and increasing productivity. By integrating multimodal learning approaches, we aim to enhance the capabilities of robotic vision systems, enabling them to handle diverse and dynamic real-world environments with greater precision and reliability. The following section reviews related work in this domain, highlighting the challenges and existing solutions that inform our research.

2.10.1 Multimodal Large Language Models for Robotic Vision

Recently, Multimodal Large Language Models (MLLMs) have achieved new state-of-the-art result, delivering superior performance across a wide range of vision and vision-language benchmarks. Their exceptional capabilities are most notable in few-shot and zero-shot scenarios, as evidenced by a series of studies [3, 50, 123, 252, 275]. Their efficacy comes from extensive pretraining on large-scale corpora of image-text data, enabling them with superior transfer capabilities. Palm-E, a pioneer in the use of MLLMs in the robotics domain, has exhibited state-of-the-art performance in embodied robotic planning scenarios [50]. Furthermore, research indicates that the multimodal large-scale training of image-text pairs results in models with enhanced robustness to out-of-distribution examples [154] compared to their unimodal counterparts. This enables MLLMs show superior performance in unimodal tasks, such as vision tasks, when compared to vision-specific pretrained models [123, 252]. As such, MLLMs present a compelling case for serving as backbone encoders in complex robotic vision applications, which often consist of multiple sub-tasks. In the method we propose in Chapter 8, we take the advantage of the power of MLLMs and tune the method for robotic vision tasks. This enables us to improve the effectiveness of robotic vision pipeline yet significantly reduce the engineering efforts and tuning time.

2.10.2 Image Segmentation in Robotic Vision

Object instance segmentation involves simultaneously predicting pixel-level instance masks and their corresponding class labels. Though the most prevalent backbones for these detector models have been Convolutional Neural Networks (ConvNets) [118], such as R-CNN [66] and Faster R-CNN [201], the Vision Transformer (ViT) [49] has emerged as a potent alternative for image classification tasks. The original ViT architecture is non-hierarchical; this characteristic hinders its applicability in object instance segmentation due to a lack of innate translational equivariance [48] and difficulties in handling high-resolution inputs because of the quadratic complexity of self-attention. Therefore, some models aim to mitigate these challenges by incorporating ConvNet designs into ViT, integrating hierarchical structures and translation-equivariant priors such as convolutions, pooling, and sliding windows (e.g., Swin [140], MViT [57], PVT [251]). These approaches compromise the model's general applicability by coupling pre-training and fine-tuning requirements. Subsequent efforts have explored plain ViT backbones specifically for object instance segmentation, such as UViT [32], a single-scale Transformer for object detection. Unlike UViT, ViTDet [130] offers an approach that retains the task-agnostic nature of ViT backbones, thereby facilitating their broader applicability. In Chapter 8 of this thesis, we extend the line of inquiry initiated by ViTDet in the robotic vision domain, opting for a plain backbone architecture decoupled from the detection task, which allows for easier deepening the intra-modal alignments.

2.10.3 Image Retrieval in Robotic Vision

ARMBench is a large-scale benchmark dataset designed for perception and manipulation challenges in robotic pick-and-place settings. Collected in an Amazon warehouse, it captures a wide variety of objects and configurations. The dataset includes images and videos of different stages of robotic manipulation, such as picking, transferring, and placing, all accompanied by high-quality annotations. Object identification in ARMBench can be tackled as an image retrieval task, a well-studied area in computer vision with applications in robotics for scene localization [4] and place recognition [35]. Traditional methods have focused on aggregating ConvNet feature maps using various pooling techniques such as R-MAC [237] and GeM [188], the latter of which has achieved state-of-the-art performance. Recently, the advent of vision transformers [49] has initiated a shift away from ConvNets, often surpassing them in performance [54, 218] and reducing the need for specialized aggregation methods [54]. As for the objective functions for training, they have commonly employed contrastive or triplet loss functions [38, 68, 157]. Thus, contrastive learning has been shown to be effective in training image retrieval tasks.

In Chapter 8 of this thesis, we fine-tune a Multimodal Large Language Model (BEiT-3) with contrastive loss, leveraging its large-scale pretraining for effective object identification. We argue that if the produced embeddings are sufficiently good, there is no need for complex feature processing methods or ranking techniques. This aligns well with the objective of RoboLLM, which aims to reduce engineering efforts in model tuning and enhance efficiency.

2.10.4 Defect Detection in Robotic Vision

Research in defect detection has been primarily oriented toward identifying surface defects in materials such as fabric, metals, and concrete [23, 152, 195]. Though the primary goal is to ascertain the existence of a defect, certain applications necessitate the specific type of defect to be classified, localized, and segmented [121, 230]. Feng et al. [60] employed an autoencoder pretraining method for defect detection with limited training data, while Hu et al. [87] introduced a lightweight spatial-temporal model incorporating local attention and PCA reduction to detect thermography defects. In contrast, ARMBench presents a large-scale challenge in which defects can manifest in various forms. Obtaining examples of defects for all objects is often infeasible, necessitating a model that can generalize to defects in unseen forms. We hypothesize that MLLMs, which are pretrained on large-scale datasets, can significantly improve the model's ability to handle out-of-distribution samples.

2.11 Cross-modal Retrieval

In Chapter 9 of this thesis, we address the third application: text-to-image retrieval. Text-to-image retrieval aims to locate relevant images in a database given a text query, which has a wide range of use-cases such as digital libraries, e-commerce, and multimedia databases. Consequently, there is a growing interest in developing effective models for this task.

The adoption of Convolutional Neural Networks (CNNs) into image retrieval marked a transformation in feature extraction capabilities. Key architectures such as VGG [214] and ResNet [81] paved the way for efficient representation of visual data. On the textual front, Recurrent Neural Networks (RNNs) and their offshoots, LSTMs and GRUs, offered significant advancement in processing sequences, rendering them suitable for textual data [158]. From around 2015, efforts turned towards creating shared embeddings for both images and text. One noteworthy method was the Deep Visual-Semantic Embedding model (DeViSE) [61], which merged visual and textual information in a common vector space. A multimodal residual network was proposed by Wang et al. [247], achieving state-of-the-art results on several benchmark datasets. In recent years, the community shift the research interest to transfer Learning and Pre-trained Architectures. The success of pre-trained models in NLP, especially the transformer architecture like BERT [46], prompted exploration in image-text retrieval. The visual counterpart, ViT (Vision Transformer), was introduced, which treats images as sequences of patches [49]. Cross-modal pre-trained models, such as UNITER [34], combined images and text in a unified framework, learning shared representations that drastically improved retrieval performance. Zero-shot is also a critical evaluation aspect of VL models . Therefore, with the rise of Few-shot Learning, models like CLIP [189] exemplified the capacity for zero-shot transfer across a range of visual and language tasks, indicating the growing unity between image and text models. Few-shot learning methods, like TIRG [102], explored using minimal data for effective image-text retrieval, crucial for real-world applications where abundant labelled data isn't always available.

Nevertheless, there are still many challenges in this area, for example:

- Image-text retrieval remains computationally intensive, with efforts to streamline and optimize current models [135].

- Exploring the synergy between unsupervised learning and retrieval tasks offers promising results [31].

2.11.1 Small-scale and Caption-based Text-to-Image Retrieval

A majority of existing Text-to-Image retrieval methods [26, 65, 93, 126, 141] concentrate on small-scale and caption-based benchmarks, such as MSCOCO [133] and Flickr30K [274]. They often excel by employing intra-modal and inter-modal attention mechanisms to align entity semantics across modalities. Specifically, they try to ensure that the meaning or representation of specific entities is consistent when interpreted through different modalities. Methods such as [59, 187, 208, 249, 250] adopt a two-stage retrieval strategy to further refine the feedback results based on one-stage ranking to obtain more accurate retrieval. For instance, MTFN [250] introduced a generic text-to-image re-ranking scheme for refinement during the inference process without requiring additional training procedures. Moreover, JGCAR [249] and LeaPRR [187] proposed modeling the higher-order neighbor relationship-aware attentions for text-image retrieval in a learnable two-stage re-ranking paradigm. However, most of these methods designed a relatively complex first-stage multimodal interaction model to establish a precise candidate set for the subsequent stage re-ranking process. These methods are computationally intensive and do not scale well to large datasets with long textual queries and diverse topics of images. Recent advancements in Multimodal Large Language Models (MLLMs) [36, 124, 132, 189, 216, 234, 252, 288] signify a paradigm shift in the field. While these models offer robust performance in Text-to-Image retrieval, their application to large-scale, long-text queries for image retrieval presents challenges in both efficiency and effectiveness due to the computational cost of MLLMs and issues with injective embeddings [221, 266]. In response, in Chapter 9, we proposes a novel two-stage coarse-to-fine index-shared retrieval framework tailored to address these challenges.

2.11.2 Large-scale Text-to-Image Retrieval datasets

AToMiC [266] is a recently released dataset for large-scale long-text to image retrieval, introduced by TREC [2]. AToMiC is built upon the WIT dataset [266]. AToMiC distinguishes itself by focusing on section-level image-text associations for multimedia content creation, emphasizing the use of English Wikipedia sections without images for a more realistic text-image context.

[2] https://trec.nist.gov

Unlike WIT, which is employed for broader tasks like image-caption matching and generation, AToMiC is tailored for ad hoc retrieval tasks, reusing WIT's images and metadata but providing image pixel values in a standardized format. Therefore, AToMiC is the only and the best option to evaluate the performance of models in the context of large-scale long-text to image retrieval. evaluate long In this paper, we target the large-scale long-text to image retrieval (LLIR) task in AToMiC. This task is designed to retrieve images from large image collections based on a long-text query for scenarios such as article writing. It encompasses more than 21 million images and textual documents, offering two distinct evaluation settings: a base setting and a large setting.

Both settings utilize a training set comprising 4,401,903 query-document relevance assessments (qrels), a validation set with 17,801 qrels, and a test set containing 9,873 qrels. In the base setting, each image is accompanied by at least one corresponding long document, and retrieval candidates are limited to the labelled images that are as least relevant to one document. In contrast, the large setting extends the candidate pool by incorporating an additional 7,608,283 images, offering a more challenging retrieval context.

In Chapter 9, we use the AToMiC dataset to train and evaluate the performance of our proposed framework.

Challenges in MLLM-based approaches In contrast to small-scale image-caption datasets such as MS COCO [133], as shown in Figure 9.2, which comprises 165,000 images with text descriptions averaging 11.53 tokens and corresponding to a single ground-truth image, AToMiC presents a more realistic simulation of LLIR applications. It uses longer, multi-faceted (ambiguous) real-world documents, averaging 415 text tokens, and maps them to multiple ground-truth images. These characteristics of LLIR introduce challenges to state-of-the-art MLLM-based approaches, which are primarily in the realms of retrieval effectiveness and computational efficiency.

In regard to retrieval effectiveness, the issues manifest in two ways. First, the complexity and multi-faceted nature of long documents introduce semantic ambiguities, making it more difficult for injective models to accurately discern text-to-image similarities. Second, this challenge is further exacerbated by the expanded pool of candidate images, complicating the task of identifying the most relevant matches.

On the computational front, inefficiencies are also divided into two parts. Firstly, the inference stage in current MLLM-based methods demands an exhaustive pairing of each query with database items, which are then processed by the MLLM to predict matching scores [122, 127, 129, 151, 231]. This model-based similarity inference is both computationally intensive and time-consuming, particularly when compared to vector-based distance computations. Secondly, the act of encoding long documents for semantic matching is itself a time-consuming process. These inefficiencies limit the practical applicability of MLLMs to large-scale retrieval tasks, despite their promising accuracy. Therefore, in Chapter 9, we propose an efficient multimodal learning framework to tackle these issues.

2.12 Keywords and Definitions

Thus far, we have presented a comprehensive overview of the background and relevant literature. This section will outline additional definitions for some key terms commonly used in this thesis to enhance your understanding.

Alignment

Alignment refers to the process of ensuring that information from different modalities (e.g., text, images, audio) corresponds accurately to the same underlying concept or meaning [78, 131]. In the context of multimodal systems, alignment often involves mapping features from different modalities into a shared or comparable representation space, enabling effective integration and analysis. Alignment can be shallow or deep, depending on the level of integration, as detailed below.

Shallow Alignment

Shallow alignment involves aligning modalities at a lower or less abstract level, typically by synchronizing features from different modalities without deep semantic integration. For instance, shallow alignment may involve concatenating or aligning features based on simple spatial or temporal correspondences [78, 131]. This method is computationally efficient and suitable for tasks where high-level semantic understanding is not crucial.

Deep Alignment

Deep alignment, in contrast, integrates modalities at a higher, more abstract level. It involves learning complex, shared representations that capture the deep semantic relationships between modalities [78, 131]. This approach is essential for tasks requiring nuanced understanding and interaction between modalities, such as multimodal reasoning or cross-modal retrieval.

Increased Alignment

Increased alignment refers to the enhancement of correspondence and coherence between multiple modalities or features in a multimodal system [12, 220]. This improvement can occur through refined techniques for mapping, integrating, or synchronizing data from different modalities. Increased alignment often involves optimizing representations to better capture semantic, temporal, or spatial relationships, thereby reducing ambiguity or misinterpretation. The goal of increased alignment is to ensure that the modalities complement each other more effectively, leading to improved performance in downstream tasks such as multimodal reasoning, cross-modal retrieval, and predictive modeling.

Measuring the Increase in Alignment

The increase in alignment between modalities can be quantified through various metrics and evaluation methods, depending on the task and the system architecture. Common approaches [12, 220] include:

1. Representation Similarity Metrics

To measure the alignment in a shared representation space, cosine similarity is often used. **Cosine Similarity:** Quantifies the similarity between feature vectors of different modalities in the shared space. Increased alignment results in higher cosine similarity values for semantically related data.

2. Cross-Modal Retrieval Accuracy

In tasks such as image-to-text or text-to-image retrieval, increased alignment leads to higher retrieval accuracy. Metrics include:

- **Precision@K:** Evaluates the proportion of correctly retrieved items in the top-K results.
- **Recall@K:** Measures the proportion of relevant items retrieved within the top-K results.
- **Mean Reciprocal Rank (MRR):** Averages the reciprocal ranks of correct results over multiple queries.

3. Downstream Task Performance

The effectiveness of alignment can also be evaluated indirectly by assessing improvements in the performance of downstream tasks such as:

- **Multimodal Classification:** Increased alignment often results in higher classification accuracy.
- **Multimodal Reasoning:** Multimodal tasks typically benefit from enhanced alignment, reflected in improved metrics like F1 score or BLEU score.

4. Alignment Loss Reduction

Many alignment models include a loss term specifically designed to encourage cross-modal coherence, such as:

- **Contrastive Loss:** Reduces the distance between representations of related data while increasing the distance for unrelated pairs.
- **Triplet Loss:** Optimizes alignment by ensuring that positive pairs are closer than negative pairs by a certain margin.

A reduction in these losses over training epochs indicates increased alignment.

Mixture of Modality Experts

A *mixture of modality experts* is a computational strategy used to model and integrate information from multiple modalities [252]. Each modality expert specializes in processing data from a specific modality (e.g., vision, language, or audio) and contributes its outputs to a unified decision-making process. This approach allows the system to leverage the unique strengths of each modality while mitigating the effects of noisy or irrelevant information in any single modality.

The combination of these concepts—alignment, modality experts, shallow alignment, and deep alignment—forms the foundation for advanced multimodal learning systems. These systems aim to effectively process and integrate diverse types of data to achieve superior performance in complex tasks.

Chapter 3

Benchmarks and Datasets

In this chapter, we provide a comprehensive overview of the datasets and benchmarks utilized in this thesis, divided into two main sections. We demonstrate the main features and statistic of these datasets and benchmarks in Table 3.1. The first section, 3.1 , covers vision training datasets and benchmarks, including widely recognized datasets such as ImageNet, CIFAR-10, and other commonly used datasets. These are critical for training and evaluating the visual components of our models. The second section , 3.2, focuses on multimodal learning training datasets and benchmarks, featuring datasets like MS COCO and Conceptual Captions. These datasets are essential for developing and assessing our multimodal learning framework, enabling the integration and alignment of diverse data types to enhance performance across various applications. By detailing these resources, we aim to underscore their significance in the development and validation of our proposed solutions.

Characteristic	Dataset	Main Features	Number of Samples	Number of Classes
Vision	ImageNet	Large-scale visual recognition, diverse categories	14M images	1,000
	CIFAR-10	Small image dataset	60K images	10
	CIFAR-100	Small image dataset	60K images	100
	Caltech-256	Large number of object categories	30K images	256
	Oxford-Flowers	Flower species classification	8K images	102
	Oxford-IIIT Pet	Pet species classification	7.4K images	37
	iNaturalist	Biodiversity, fine-grained categories	437K images	10K+
	Places365	Scene recognition	1.8M images	365
Multimodal	VQA	Visual question answering	204K images, 1.1M questions	-
	NLVR	Natural language for visual reasoning	74K images, 92K sentences	-
	MS COCO	Common objects in context, captions	330K images, 5 captions/image	-
	Flick30k	Image captioning	31K images, 158K captions	-
	GQA	Balanced question answering	113K images, 22M questions	-
	Conceptual Captions	Image captioning from web	3.3M images, captions	-
	SBU Captioned Photo Dataset	Captioned photos from Flickr	1M images, captions	-

Table 3.1: Description of Vision Training Datasets and Multimodal Training Datasets and Benchmarks

3.1 Vision Training Datasets and Benchmarks

In this section, we provide a comprehensive overview of the vision training datasets and benchmarks utilized in our research. The inclusion of these datasets and benchmarks is crucial as they form the foundation for training and evaluating the visual components of our multimodal learning framework. By leveraging diverse and representative datasets, we ensure that our vision encoders are robust and capable of handling real-world scenarios. These datasets not only facilitate the development of effective image models but also play a pivotal role in enhancing the overall performance of our multimodal framework. Understanding the characteristics and challenges of these benchmarks allows us to fine-tune our models for optimal intra-modal and inter-modal alignment, thereby improving the efficacy of our proposed solutions across various applications, including crisis response, cross-modal retrieval, and recommendation systems.

3.1.1 ImageNet

The ImageNet dataset is a landmark resource in the field of computer vision, playing a critical role in advancing technologies related to image classification, object recognition, and deep learning. Developed by Deng et al. [45], this extensive dataset was designed to mirror the structure of the human visual wordnet, making it both broad and detailed in scope. It contains more than 14 million labeled images collected from the web, each categorized into over 20,000 categories, providing a diverse and comprehensive visual representation of objects and scenes.

One of the most influential aspects of ImageNet is its use in the annual ImageNet Large Scale Visual Recognition Challenge (ILSVRC), which has been a key driver in the development of advanced neural network architectures. The challenge tasks participants with classifying images into thousands of categories, detecting objects, and localizing them within the image. It has been a proving ground for pioneering architectures like AlexNet [114], ResNet [81], and VGG [215], each of which has set new benchmarks in accuracy and efficiency.

ImageNet's depth and complexity challenge algorithms to develop robust and accurate models capable of handling real-world variability in object appearance, pose, and lighting conditions. The dataset has not only been instrumental in the evolution of computer vision but has also impacted other fields by facilitating advancements in machine learning techniques that leverage large-scale data. As such, ImageNet remains a foundational tool for researchers and developers aiming to create more intelligent and adaptive systems in the ever-evolving landscape of AI technology.

3.1.2 CIFAR-10 and CIFAR-100

The CIFAR-10 and CIFAR-100 datasets are fundamental resources in the field of machine learning, specifically designed for training and testing image recognition systems. Developed by

Krizhevsky et al. [112], these datasets provide a compact yet challenging benchmark for evaluating algorithmic performance in visual classification tasks.

The CIFAR-10 dataset consists of 60,000 32x32 color images divided into 10 classes, with 6,000 images per class. The dataset is split into 50,000 training images and 10,000 testing images, encompassing a variety of everyday objects such as airplanes, cars, birds, and cats. This dataset is particularly notable for its balance across classes, providing a uniform challenge for machine learning models.

Expanding on the CIFAR-10, the CIFAR-100 dataset is structured similarly in terms of image size and number but offers a finer classification with 100 classes, each containing 600 images. These classes are grouped into 20 superclasses, adding an additional layer of hierarchical labeling that can be used for more nuanced learning and classification tasks. Like CIFAR-10, the CIFAR-100 is also divided into 50,000 training images and 10,000 testing images.

Both datasets are derived from the '80 million tiny images' dataset and are intentionally designed to be computationally manageable while still offering a significant challenge due to the low resolution and high variability of the images. The CIFAR datasets have been instrumental in driving advances in computer vision, particularly in the development and refinement of convolutional neural networks (CNNs), by providing a standard, reproducible benchmark for researchers to evaluate new models, algorithms, and techniques in image classification.

3.1.3 Caltech-256

The Caltech-256 dataset is a prominent image dataset in the field of computer vision, specifically designed to further the development of object recognition systems. Developed by Griffin et al. [72], it serves as an extension and enhancement of the earlier Caltech-101 dataset, offering a more challenging testbed for machine learning models due to its larger number of categories and images.

Caltech-256 contains a total of 30,607 images categorized into 256 object classes, plus a background/clutter category. Each class has at least 80 images, which significantly improves the dataset's utility for training robust object recognition models. The images in Caltech-256 are more diverse and include a greater degree of intra-class variability and background clutter than its predecessor, Caltech-101, providing a more rigorous challenge that better simulates real-world conditions.

The dataset encompasses a wide range of objects from various everyday categories, including animals, vehicles, household objects, and scenic images. This variety, coupled with the challenging nature of the images, tests the limits of existing algorithms and drives innovation in areas such as feature extraction, object classification, and the generalization capabilities of computer vision systems.

Caltech-256 has been instrumental in advancing the field of object recognition. It provides a comprehensive platform for developing and benchmarking algorithms, particularly those employing techniques such as convolutional neural networks (CNNs) and other advanced machine learning methods aimed at handling complex visual data. The dataset's impact is reflected in its widespread use across academic research and its contribution to the improvement of practical applications in image-based recognition and classification systems.

3.1.4 Oxford-Flowers

The Oxford-Flowers dataset is a specialized dataset designed for image classification tasks, particularly focusing on flower species recognition. Developed by Nilsback and Zisserman [168], this dataset is part of the broader effort to enhance computer vision systems' ability to recognize natural objects within uncontrolled environments.

The Oxford-Flowers dataset includes two key collections: the Oxford-Flowers 17 and Oxford-Flowers 102. The Oxford-Flowers 17 contains 1,360 images divided into 17 different flower classes, each class corresponding to a species commonly found in the United Kingdom. Each class has 80 images for training, and the images are taken under natural conditions with variations in scale, lighting, and viewpoint.

Expanding significantly on this, the Oxford-Flowers 102 consists of 8,189 images representing 102 flower species, encompassing a more comprehensive range of common flowers in the UK. This dataset is challenging due to the high variability in appearance of the flowers due to natural factors such as growth stage, occlusion, and environmental conditions.

The dataset images were acquired from various sources, including web searches and photographs taken by the researchers, ensuring a realistic variation in image quality and background complexity. Each image in the datasets is annotated with the corresponding flower class, which facilitates supervised learning tasks.

These datasets are particularly useful for advancing algorithms in fine-grained visual categorization, a sub-field of computer vision that focuses on distinguishing between highly similar objects. The Oxford-Flowers datasets have been instrumental in developing and benchmarking algorithms for image classification, and they continue to support new approaches in deep learning, feature extraction, and more robust model training in the context of natural image settings.

3.1.5 Oxford-IIIT Pet

The Oxford-IIIT Pet dataset is an enriched image dataset designed specifically for fine-grained visual categorization, focusing on pet animals. Developed by Parkhi et al. [179], this dataset provides a rich test bed for algorithms aiming to distinguish between different breeds of cats and dogs, a challenging task given the subtle differences between breeds.

The dataset comprises 7,349 images representing 37 different pet breeds (12 cat breeds and 25 dog breeds), with approximately 200 images per breed. Each image has been meticulously annotated to include not only the breed labels but also head annotations, with each image tagged with a head region and a tight bounding box around the pet. These annotations are critical for tasks that require understanding of the structure and features specific to each breed, such as facial recognition technologies and advanced classification systems.

One of the key challenges posed by the Oxford-IIIT Pet dataset is the inherent variability in the images, which include differences in pose, scale, lighting, and background. These factors mimic real-world conditions, thus providing a realistic scenario for developing robust algorithms capable of accurate breed classification.

The Oxford-IIIT Pet dataset has become a valuable resource for researchers in the field of computer vision, particularly for those working on fine-grained classification tasks. It has also been used extensively to benchmark and refine the performance of various image recognition models, including those based on convolutional neural networks (CNNs), which benefit from the detailed annotations and high-quality images provided in the dataset. This dataset continues to support advancements in image processing techniques and contributes to the improvement of practical applications involving pet identification and animal welfare technologies.

3.1.6 iNaturalist

The iNaturalist 2017 (iNat2017) dataset [240] is a specialized resource aimed at fostering advancements in fine-grained visual categorization, particularly focusing on species identification. Developed as part of the iNaturalist species classification and detection competition at the FGVC4 workshop (held in conjunction with CVPR 2017), this dataset presents a unique set of challenges due to the natural variability in species appearance.

The iNat2017 dataset features over 675,000 images representing approximately 5,000 species, ranging from plants and insects to birds and mammals. The dataset is derived from observations submitted by citizen scientists around the world through the iNaturalist platform, making it one of the largest and most diverse collections of natural images available for research.

Each image in the dataset is annotated with species labels, and many include additional information such as the location and time of the observation. This contextual data can be invaluable for tasks that require understanding environmental influences on species appearance or behavior. The dataset's diversity and size challenge existing image recognition systems to improve their accuracy and robustness, particularly in conditions of variable lighting, occlusion, and background.

The iNat2017 dataset is not only a tool for advancing machine learning techniques but also serves as a bridge between technology and biodiversity conservation. By improving the ability of algorithms to recognize and classify species from photographs, this dataset supports efforts to monitor biodiversity, track species distributions, and engage the public in scientific research. It remains a critical resource for researchers in both the fields of computer vision and conservation biology, promoting deeper understanding and protection of the natural world.

3.1.7 Places365

The Places365 dataset is an extensive image collection designed specifically for scene recognition tasks, an area of computer vision focused on identifying and classifying different environments rather than objects. Developed by Zhou et al. [290], this dataset is part of the larger Places project, which aims to provide a comprehensive resource for training models to understand the context and settings of various scenes.

Places365 contains over 1.8 million images across 365 scene categories, representing a diverse array of indoor and outdoor environments. Categories range from common rooms like kitchens and living rooms to outdoor landscapes such as beaches, fields, and urban scenes. This wide variety helps ensure that models trained on the dataset can generalize across a variety of real-world scenarios.

Each image in the dataset is carefully annotated with its scene category, facilitating the development of robust algorithms capable of distinguishing subtle differences between similar categories, such as different types of restaurants or natural landscapes. The dataset is designed to challenge and enhance the performance of deep learning models, particularly those using convolutional neural networks (CNNs), in recognizing and categorizing complex scenes.

The Places365 dataset is crucial for advancing scene recognition technologies, which have wide-ranging applications including robotics navigation, augmented reality, and automated tagging in photo libraries. By providing a rich, varied collection of scene images, Places365 allows researchers and developers to push the boundaries of what AI systems can understand about the visual world, improving how machines interpret and interact with their surroundings.

3.1.8 Summary

The aforementioned datasets will primarily be used for developing contrastive learning methods in Chapter 5 and for training the image encoder in the proposed multimodal learning framework, MCA, detailed in Chapter 6.

3.2 Multimodal Training Datasets and Benchmarks

In this section, we provide a detailed overview of the multimodal training datasets and benchmarks that are integral to our research. These datasets and benchmarks are crucial for the development and evaluation of our proposed multimodal learning framework. By incorporating a variety of data types, such as text, images, and other modalities, these multimodal datasets enable the comprehensive training of models that can effectively integrate and align information from different sources. This alignment is essential for enhancing the performance and robustness of our framework across diverse applications. The benchmarks serve as standardized measures to assess the efficacy of our models, ensuring that they perform well not only in controlled environments but also in real-world scenarios. By understanding and leveraging these datasets, we can refine our framework to achieve superior intra-modal and inter-modal alignment, thereby optimizing its application in areas such as crisis response, cross-modal retrieval, robotics, and recommendation systems.

3.2.1 VQA

The Visual Question Answering (VQA) dataset is a pivotal resource designed to advance the field of machine learning, specifically focusing on the integration of visual and textual information. Introduced by Antol et al. [5], this dataset facilitates the development and evaluation of models capable of answering open-ended questions about images. Central to its structure is a collection of images paired with a series of questions that challenge the model's understanding of visual content in relation to textual queries. Each question is accompanied by multiple answers, which are typically obtained through crowd-sourcing, ensuring a diverse representation of human perception and opinion. This dataset not only supports the enhancement of algorithms for image understanding and natural language processing but also propels research in multimodal learning, where the intersection of different types of data modalities is crucial. The VQA dataset, therefore, serves as a fundamental tool for researchers and practitioners aiming to build systems that better mimic human-level understanding in complex, multimodal environments. The challenges of the VQA task stem from the ambiguity of certain questions and answers, as well as question bias.

3.2.2 NLVR

The Natural Language for Visual Reasoning (NLVR) dataset is an essential resource developed to enhance the capabilities of machine learning systems in the area of visual reasoning. Introduced by Suhr et al. [227], the dataset is designed to assess the ability of models to interpret and reason about images based on structured textual descriptions. Unlike simpler image captioning tasks, NLVR requires a model to verify the truth of a statement given one or more images, adding a layer of complexity that involves logical deduction and deeper semantic understanding. This dataset includes a diverse set of images paired with statements that are either true or false, meticulously annotated to challenge the model's reasoning abilities. The NLVR dataset supports advancements in the field by pushing the boundaries of how AI systems integrate and interpret multimodal information, aiming to achieve a more nuanced and accurate understanding of visual content in the context of natural language descriptions.

3.2.3 MS COCO

The Microsoft Common Objects in Context (MS COCO) dataset is a cornerstone in the fields of computer vision and machine learning, tailored to enhance technologies in image recognition, segmentation, and captioning. Introduced by Lin et al. [133], this dataset features a rich collection of images that depict complex everyday scenes with common objects in natural contexts.

MS COCO is renowned for its detailed annotations, which include object bounding boxes, segmentation masks, and comprehensive captions, facilitating precise object localization and extensive scene understanding. The dataset encompasses over 330K images, with more than 200K labeled for training. It covers 91 object types that have been annotated with over 1.5 million object instances, making it one of the most extensive datasets available for object detection and image captioning.

This extensive labeled dataset is structured to challenge and refine algorithms by presenting varied and complex scenarios that demand both detailed recognition capabilities and a broader contextual understanding of scenes. Consequently, MS COCO has become an indispensable resource for researchers and practitioners aiming to advance the perceptual abilities of AI systems, significantly contributing to the development of technologies that allow machines to interpret and describe visual information both accurately and contextually.

3.2.4 Flick30k

The Flickr30k dataset is a prominent resource in the field of computer vision and natural language processing, specifically designed to enhance the capabilities of image captioning systems. Introduced by Young et al. [274], this dataset comprises 31,000 images sourced from the online photo-sharing platform, Flickr. Each image in Flickr30k is accompanied by five detailed captions, which are independently written by human annotators. This structure enables the development and evaluation of algorithms that must understand and describe complex visual scenes in natural language.

Flickr30k offers a diverse array of everyday scenes and events, capturing a wide range of human activities, objects, and settings that are commonly encountered in daily life. The inclusion of multiple captions per image not only provides a rich set of linguistic expressions describing the same visual content but also encapsulates varied interpretations and descriptive focuses, enhancing the training of more robust and versatile models.

This dataset is particularly valuable for training and benchmarking models in tasks such as automatic image captioning, visual question answering (VQA), and multimodal machine translation. By bridging the gap between visual data and natural language, Flickr30k helps advance the development of AI systems capable of understanding and generating human-like descriptions of visual content, thereby contributing significantly to the field's progress towards more sophisticated multimodal interactions.

3.2.5 GQA

The GQA dataset is a specialized resource developed to advance research in visual reasoning and natural language understanding within the context of question answering systems. Introduced by Hudson and Manning [95], the GQA dataset offers a structured environment for evaluating and enhancing machine learning models' abilities to parse and reason about complex visual scenes using a question-and-answer format.

This dataset consists of over 22 million question-answer pairs, which are systematically generated from 113,000 images. These questions are not merely descriptive but are designed to test various levels of reasoning, including spatial understanding, relational reasoning, and logical inference. Each question is carefully crafted to reflect a balanced mix of compositionality, requiring multiple steps of reasoning, and diversity in both question type and difficulty.

The GQA dataset's unique contribution lies in its emphasis on consistency and balance, addressing common biases that often plague visual question answering datasets. It provides detailed annotations for both questions and answers, including the types of reasoning needed and the relationships among objects within the images. Furthermore, GQA supports a variety of tasks such as object detection, attribute classification, and spatial reasoning, making it a comprehensive tool for developing more sophisticated, context-aware AI models.

By focusing on structured, real-world visual reasoning, the GQA dataset plays a crucial role in pushing the boundaries of what AI systems can achieve in understanding and interacting with the visual world, thus propelling forward the capabilities of AI in interpreting complex multimodal data.

3.2.6 Conceptual Captions

The Conceptual Captions dataset is a large-scale resource designed to bolster advancements in image captioning and the broader field of multimodal machine learning. Developed by Sharma et al. [210], this dataset provides a substantial collection of image-caption pairs, which are uniquely generated by harvesting and transforming alt-text data from the web into image captions. The transformation process involves removing any web-specific context to make the captions more generalizable and applicable to a broader range of images.

Comprising over 3.3 million image-caption pairs, the Conceptual Captions dataset is characterized by its diversity and complexity of both visual content and associated textual descriptions. Unlike other datasets where captions are often manually annotated and may follow a more consistent format, the captions in Conceptual Captions are sourced from a variety of online contexts, resulting in a wide range of linguistic expressions and styles. This variability makes it an excellent resource for training models to understand and generate natural language descriptions of images in a way that is less constrained and more reflective of how people naturally describe scenes.

The dataset is particularly valuable for training deep learning models in tasks such as automatic image captioning, visual question answering, and other AI-driven applications where the interaction between visual data and natural language is crucial. By providing a bridge between these two modalities, the Conceptual Captions dataset helps in developing more sophisticated, context-aware AI systems that can operate effectively in diverse and dynamic environments.

Overall, the Conceptual Captions dataset represents a significant step forward in the creation of AI that can interpret visual content with the same richness and diversity as human language, enhancing the capabilities of systems to engage in more intuitive and meaningful multimodal interactions.

3.2.7 SBU Captioned Photo Dataset (SBU)

The SBU Captioned Photo Dataset, developed by Ordonez et al. [176], is a substantial collection specifically curated to facilitate research in automatic image captioning and vision-language integration. This dataset comprises one million image-caption pairs, sourced primarily from Flickr, which provides a rich basis for training and evaluating machine learning models that handle both visual and textual data.

Each image in the SBU dataset is paired with a natural language caption, generated by the original image uploader, offering a genuine and spontaneous description of the scene. This characteristic is particularly valuable as it captures a wide variety of human perceptions and linguistic expressions, reflecting real-world usage of language in describing visual content.

The diversity in the dataset extends not only to the captions but also to the images themselves, which depict a broad range of subjects including people, animals, objects, and landscapes in various settings and situations. This variety ensures that models trained on the SBU dataset can develop robust capabilities in understanding and generating descriptions across a wide array of scenes and contexts.

The SBU Captioned Photo Dataset is instrumental for advancements in several key areas of AI research, including but not limited to image captioning, automatic metadata generation for visual content, and the training of models for more sophisticated vision-language tasks. By providing a bridge between visual data and natural language, the SBU dataset helps to enhance the interpretative and descriptive capabilities of AI systems, making it a critical resource for developing more intuitive and context-aware AI applications.

3.2.8 Conclusions

In conclusion, the datasets and benchmarks detailed in this chapter provide a comprehensive foundation for our research. By utilizing a diverse array of vision training datasets, such as ImageNet, CIFAR-10, CIFAR-100, Caltech-256, Oxford-Flowers, Oxford-IIIT Pet, iNaturalist, and Places365, we ensure robust training and evaluation of the visual components of our multimodal learning framework. Additionally, the inclusion of multimodal training datasets like VQA, NLVR, MS COCO, Flick30k, GQA, Conceptual Captions, and the SBU Captioned Photo Dataset allows us to effectively integrate and align various data modalities. This diverse dataset collection supports a wide range of tasks and is instrumental in validating our experiments and substantiating our claims. By leveraging these datasets, we can demonstrate the efficacy and versatility of our proposed multimodal learning framework across multiple applications, thereby reinforcing the contributions and findings of this thesis.

Chapter 4

The Proposed MCA Framework

In this chapter, we introduce the MCA framework based on new Mixture-of-Modality-Experts (MoME) design for multi-modal model learning, as shown in Figure 4.1, that aims to enhance intra-modal and inter-modal alignment, as stated in the thesis statement (1.2). This framework forms the foundation of the thesis, as if we are able to demonstrate that the use of this framework results in both a) measurable increases in intra-modal and inter-modal alignment; and b) increased on-task effectiveness, we provide evidence supporting the thesis statement. Indeed, we instantiate this framework with a range of modular components and for a range of tasks in subsequent experimental chapters. Moreover, contrastive learning methods also contribute to improve the intra-modal and inter-modal alignment. More detailed information on MoME is provided in Section 4.1, while the contrastive learning methods are discussed in Chapter 5. Additionally, as highlighted in Section 2.7, the efficiency issue becomes more severe as the size of neural models continues to grow, often limiting their application to specialized downstream tasks due to computational constraints. To address this, we also propose a dedicated parameter-efficient learning method named MultiWay-Adapter to enhance inter-modal alignment. More details on the MultiWay-Adapter are provided in Chapter 6.

4.1 Mixture of Modality Experts (MoME)

The Mixture of Modality Experts (MoME) design [252], illustrated in Fig. 4.2, aims to address the issue of shallow intra-modal and inter-modal alignment in multimodal learning, as outlined in the thesis statement (Section 1.2). MoME is a novel approach implemented within a Transformer framework specifically for handling multimodal tasks involving both vision and language. Here, we give the key aspects of the MoME design to show how it enhances the intra-modal and inter-modal alignment. The MoME Transformer replaces the standard feed-forward network in Transformer blocks with a pool of modality-specific experts. There are two types of experts:

1. Vision Expert (V-FFN): Processes image-only inputs.

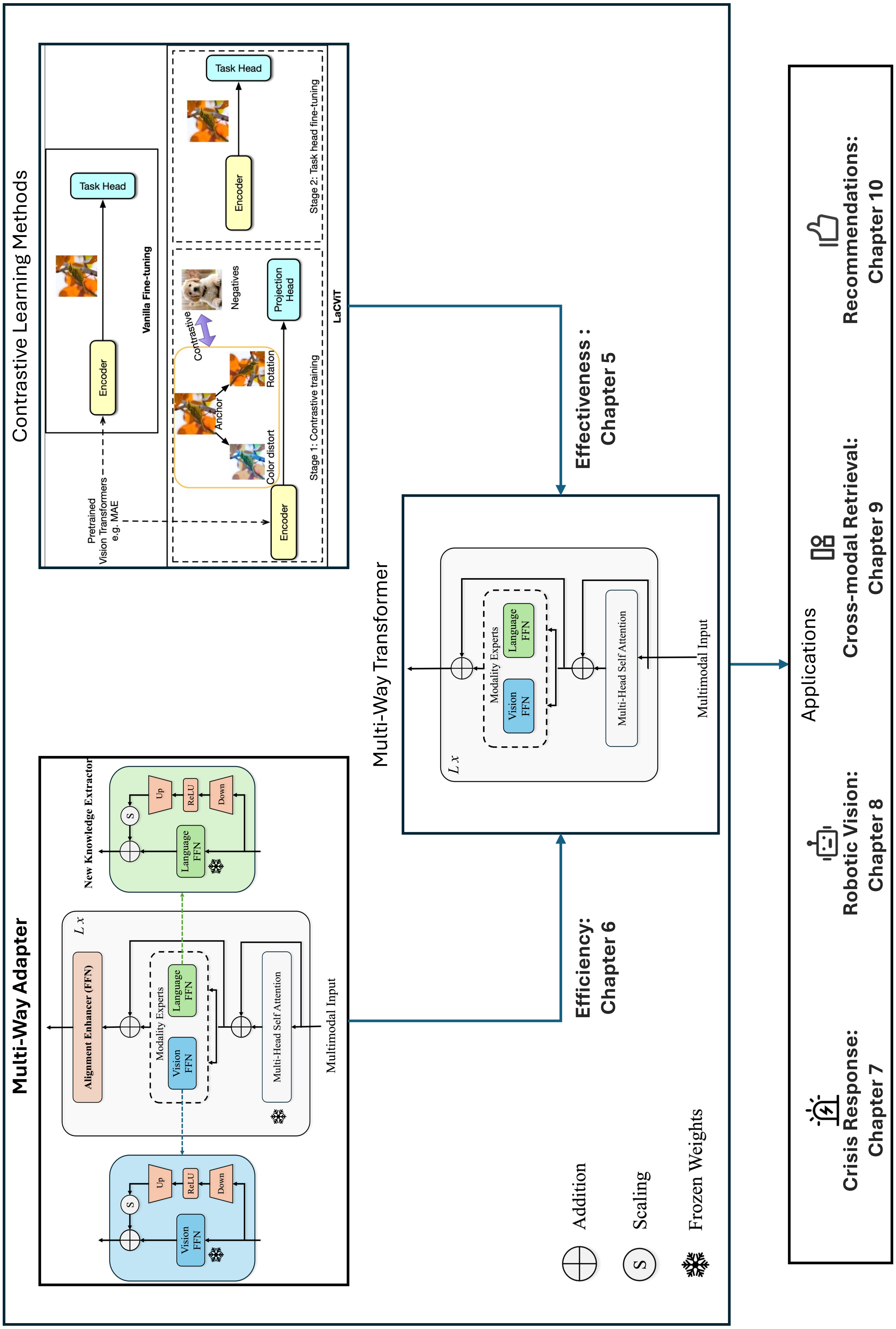

The Propose Multimodal Learning Framework: MCA
Multi-Way Adapter
Contrastive Learning Methods
Pretrained Vision Transformers e.g. MAE
Encoder
Task Head
Vanilla Fine-tuning
Anchor
Color distort
Rotation
Contrastive
Negatives
Projection Head
Stage 1: Contrastive training
Stage 2: Task head fine-tuning
LaCViT
Vision FFN
Language FFN
Up
ReLU
Down
S
New Knowledge Extractor
L x
Alignment Enhancer (FFN)
Modality Experts
Multi-Head Self Attention
Multimodal Input
Multi-Way Transformer
Effectiveness : Chapter 5
Efficiency: Chapter 6
Addition
Scaling
Frozen Weights
Applications
Crisis Response: Chapter 7
Robotic Vision: Chapter 8
Cross-modal Retrieval: Chapter 9
Recommendations: Chapter 10

Figure 4.1: The illustration of the proposed multimodal learning framework, MCA.

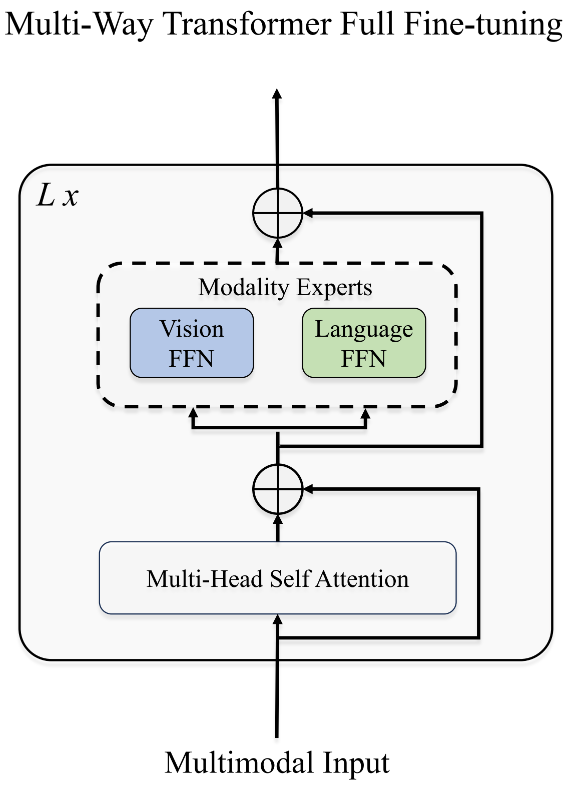

Multi-Way Transformer Full Fine-tuning
L x
Modality Experts
Vision FFN
Language FFN
Multi-Head Self Attention
Multimodal Input

Figure 4.2: The illustration of architecture of the Mixture of Modality Experts design.

2. Language Expert (L-FFN): Processes text-only inputs.

Another key design decision is the addition of shared Self-Attention: Each MoME Transformer block includes a shared multi-head self-attention layer that aligns visual and linguistic content across different modalities. This shared layer facilitates the interaction between image and text features within the same Transformer layer. This interaction deeply enhance both intra-modal and inter-modal alignment which solves the issues we identify in the thesis statement 1.2.

Flexible Modeling: Due to its flexible design, MoME can be utilized in different ways:

- Dual Encoder: For tasks like image-text retrieval, the model can encode images and text separately, allowing for efficient retrieval operations.

- Fusion Encoder: For tasks requiring deeper interaction between image and text (e.g., visual question answering), the model encodes image-text pairs together to capture complex relationships.

- Unified Pre-Training and Fine-Tuning: MoME is pretrained on large-scale datasets with multiple tasks, including image-text contrastive learning, image-text matching, and masked language modeling. This unified pre-training approach allows the model to generalize well across various downstream tasks, such as classification and retrieval.

- Stagewise Pre-Training: To leverage large-scale datasets effectively, the model undergoes a stagewise pre-training process. Initially, vision experts and self-attention modules are pretrained on image-only data. Subsequently, language experts are pretrained on text-only data, and finally, the model is fine-tuned with vision-language data, improving its ability to handle diverse inputs.

There are several advantages of MoME design:

- Enhanced modalities alignment: By training multiple experts to specialize in different regions of the input space, MoME models can capture more complex patterns and dependencies in the data compared to a single monolithic model, thereby enhancing the intra-modal and inter-modal alignment significantly.
- Efficiency: The MoME approach can improve computational efficiency. Since only a subset of experts is activated for each input, the model can be more efficient in terms of both computation and memory usage.
- Scalability: MoME models can scale to larger architectures by adding more experts without a proportional increase in computational cost for each input, as only a few experts are used at a time.

In summary, the MoME design plays a crucial role in our proposed multimodal framework, offering all the advantages mentioned above, particularly in enhancing shallow inter-modal alignment. The MoME design enables the model to perform efficiently and accurately across a variety of vision-language tasks, delivering state-of-the-art performance while maintaining flexibility and efficiency. We will provide more details of MoME in Chapter 9.

4.2 Multimodal Large Language Models (MLLMs) Used in this Thesis

In light of the advantages of deep alignment techniques for multimodal representation learning, as discussed in Section 2, we detail the settings in which we extend the use of MLLMs encoders for different applications. First, we describe the process of extracting features using multimodal embeddings obtained from CLIP, VLMo, and BEiT-3. Next, we outline the method for fine-tuning these MLLMs encoders on downstream datasets and illustrate the integration of these encoders with other task-specific models in an end-to-end approach.

4.2.1 Preliminaries on Multimodal Large Language Models

We first introduce how MLLMs encoders work, showing the input representations, propagation functions, and pre-training objectives of these MLLMs encoders.

Table 4.1: Notations used in this section to describe the proposed multimodal learning framework.

Symbol	Description
m	the multimodal embeddings of the token
l	the layer number
LN	the layer normalisation operation
MSA	the multi-head self-attention operation
FFN	the feed-forward network
p_{i2t}	the softmax-normalised image-to-text similarities
p_{t2i}	the softmax-normalised text-to-image similarities
N	the batch size
T	the length of the text sequence
$w_j^{(i)}$	the j-th word in the i-th text sequence
$w_{<j}^{(i)}$	the prefix of the i-th text sequence up to the j-th word

4.2.2 Input representations

In the following, we describe the input of the MLLMs encoders we use in this section:

(1) **CLIP:** For CLIP, raw images and texts are encoded into image and text vector representations. CLIP leverages the Vision Transformer (ViT) architecture [49] to process image representations by dividing the input image into non-overlapping patches, flattening them into vectors, and linearly projecting them to create patch embeddings. Text representations are generated using the GPT-2 [192] model, after tokenizing the raw text input using byte pair encoding (BPE) and adding positional embeddings.

(2) **VLMo & BEiT-3:** Similar to CLIP, raw images and texts are encoded into image and text vector representations. Image representations are created by splitting input images into patches, flattening them, and linearly projecting them to form patch embeddings. A learnable special token [I_CLS] is added to the sequence, and image input representations are computed by summing the patch embeddings, 1D position embeddings, and linear projection, respectively. Text representations are generated using BERT tokenization and WordPiece for subword units, with a start-of-sequence token ([T_CLS]) and a boundary token ([T_SEP]) added. Text input representations are generated by summing the word, position, and type embeddings.

To illustrate how the multimodal embeddings of CLIP, VLMo, and BEiT-3 can be used as initial input embeddings in downstream tasks, we use the VLMo-Base Plus model variant as an example. The resulting text embeddings for all items from the VLMo-Base Plus encoder have a shape of [input number, text_token_length+2, 544], where each input comprises the number of raw text tokens along with [CLS] and [SEP] tokens. As for image embeddings, each input has an embedding shape of [197, 544]. We obtain the input visual and text embeddings by selecting the 544-dimensional vector corresponding to the [CLS] token for each input embedding, which should encapsulate rich, high-level information in each modality.

4.2.3 Propagation functions

In the following, we describe the propagation functions of the used MLLMs encoders:

(1) **CLIP:** CLIP leverages a dual-stream architecture to encode distinct modalities, incorporating a separate stream for each modality—visual and textual—while maintaining shared multi-head self-attention layers to enable alignment and interaction between visual and linguistic content. Each stream consists of a series of transformer blocks. Hence, the propagation function is defined as follows:

$$h_i^{(l)} = \mathrm{LN}\left(h_i^{(l-1)} + \mathrm{MSA}(h_i^{(l-1)}, h_i^{(l-1)}, m_i)\right) \tag{4.1}$$

where $h_i^{(l)}$ is the hidden state of the i-th token in the l-th layer, LN and MSA are the layer normalization operation and the multi-head self-attention mechanism, respectively, and m_i is the corresponding multimodal embedding of the token.

(2) **VLMo & BEiT-3:** As unified LMM encoders, VLMo and BEiT-3 both use the MoME transformer to encode different modalities, with a mixture of modality experts substituting the feed-forward network of a standard Transformer [242]. Each MoME transformer block captures modality-specific information by switching to a different modality expert and employs multi-head self-attention (MSA) shared across modalities to align visual and linguistic content. Hence, the propagation function is defined as follows:

$$h_i^{(l)} = \mathrm{LN}\left(h_i^{(l-1)} + \mathrm{MSA}(h_i^{(l-1)}, h_i^{(l-1)}, m_i) + \mathrm{FFN}(h_i^{(l-1)}, m_i)\right) \tag{4.2}$$

where FFN is the feed-forward network. This mechanism of MoME-FFN is capable of selecting an expert among multiple modality experts to process the input according to the modality of the input vectors and the index of the Transformer layer. There are three modality experts: vision expert (V-FFN), language expert (L-FFN), and vision-language expert (VL-FFN). The choice of a given expert depends on the input modality and the layer within the transformer architecture. The contextualized representations for image-only, text-only, and image-text inputs are obtained accordingly.

4.2.4 Pre-training Objectives of the MLLMs Encoders:

In order to study the impact of fine-tuning the MLLMs encoders, we first describe their training objectives before adapting these encoders to the recommendation task:

(1) **Image-Text Contrastive (ITC) Loss:** All MLLMs encoders (CLIP, VLMo, and BEiT-3) leverage the ITC loss, which aims to encourage the model to learn a joint embedding space where the similarity between an image and its corresponding text is maximized, while the similarity between mismatched image-text pairs is minimized. ITC loss is defined as follows:

$$\mathscr{L}_{\text{ITC}} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{p_{i2t}(i)}{\sum_{j\neq i}^{N} p_{i2t}(j)} - \frac{1}{N}\sum_{i=1}^{N}\log\frac{p_{t2i}(i)}{\sum_{j\neq i}^{N} p_{t2i}(j)} \tag{4.3}$$

where N is the batch size, while $p_{i2t}(i)$ and $p_{t2i}(i)$ are the softmax-normalized image-to-text and text-to-image similarities of the i-th pair, respectively. As a result, a joint embedding space is learned using a contrastive loss, where the similarity between an image and its corresponding text encourages the encoder to generate more aligned embeddings.

(2) **Masked Language Modeling (MLM) loss:** This loss function is exclusively used by the VLMo and BEiT-3 encoders. These encoders randomly select and mask tokens in the text sequence with a 15% masking probability, as in BERT [46]. Both the VLMo and BEiT-3 encoders are trained to predict these masked tokens, using unmasked tokens and visual cues. The MLM loss is computed as follows:

$$\mathscr{L}_{\text{MLM}} = -\frac{1}{N}\sum_{i=1}^{N}\sum j = 1^{T_i}\log p(w_j^{(i)}|w_{<j}^{(i)}, m_i) \tag{4.4}$$

where N is the batch size, T_i is the length of the i-th text sequence, $w_j^{(i)}$ is the j-th word in the i-th text sequence, $w_{<j}^{(i)}$ is the prefix of the i-th text sequence up to the j-th word, and m_i is the corresponding multi-modal embedding of the i-th text sequence. As such, predicting masked tokens in the presence of a visual context enables the encoder to generate embeddings that better capture the joint representation of image and text data, leading to enhanced multimodal representation learning.

(3) **Image-Text Matching (ITM) loss:** This loss function is solely used by VLMo. VLMo uses the final hidden vector of the [T_CLS] token to represent the image-text pair, employing a cross-entropy loss for binary classification. Hard negatives are sampled from the training examples for this purpose. ITM loss is formulated as follows:

$$\mathscr{L}_{\text{ITM}} = -\frac{1}{N}\sum_{i=1}^{N}[y_i\log p(y_i = 1|\mathbf{h}_i) + (1-y_i)\log p(y_i = 0|\mathbf{h}_i)] \tag{4.5}$$

where y_i is the ground truth label (matched or unmatched) for the i-th image-text pair, and $\mathbf{h}_i$ is the final hidden vector of the [T_CLS] token representing the pair. The loss is computed using a binary cross-entropy loss function. By using a binary cross-entropy loss and hard negative mining, VLMo is expected to learn to differentiate between matched and unmatched image-text pairs. This process encourages the encoder to generate more accurate and semantically aligned multimodal embeddings, thereby improving the model's overall capability to mine relationships between image and text data.

4.2.5 Training Strategies of the MLLMs Encoders

In this subsection, we present various training paradigms we use in our multimodal learning framework as previously shown in 4, for incorporating the MLLMs encoders with task specific models. We discuss the optimization objectives used to tune both the MLLMs encoders and the task specific models. We aim to identify the best training paradigm that facilitates the effective integration of the MLLMs encoders, leveraging their strengths to enhance the performance of existing task specific models:

(1) **Two-stage training** involves first fine-tuning the MLLMs encoders on downstream tasks images and texts. This fine-tuning process is designed to enhance the adaptability and performance of the MLLMs encoders in the context of downstream tasks scenarios. Specifically, we tune both CLIP, VLMo, and BEiT-3 using the ITC loss [15, 189];

(2) **End-to-end training** typically needs the seamless integration of the MLLMs encoders so as to jointly optimize both the MLLMs encoders and the existing task specific models with the task specific loss. This integration must account for the different types of losses used by each model, such as the BPR loss [202] and the contrastive loss [259]. Notably, this end-to-end training process does not incorporate the previously mentioned pre-training losses of the MLLMs in Subsection 4.2.4, such as the ITC, MLM, and ITM losses. Algorithm 1 (below) presents an example pseudo-code of recommendation tasks for end-to-end training.

These training objectives and strategies form the foundation for achieving state-of-the-art performance with our proposed MCA multimodal learning framework. Unlike previous work, we systematically evaluate the impact of these methods on the performance of multimodal learning and optimize them specifically for the four domains targeted in this thesis. This optimization is detailed in Chapters 7 through 9.

4.3 Research Questions

This thesis focuses on addressing critical issues that limit the performance of multimodal learning in terms of effectiveness and efficiency. Specifically, we target the shallow intra-modal and inter-modal alignment problem present in previous multimodal learning frameworks. This leads to the following research questions:

- RQ 1: How do contrastive learning methods impact on modalities alignment? (Chapter 5)

Algorithm 1 End-to-end training

```
Step 1: Load data
data_loader ← DataLoader(dataset_path, raw_images, concatenated_texts(title, description, brand, categorical_info))
Step 2: Initialise and load pre-trained weights for CLIP/VLMo/BEiT-3 encoder
clip/vlmo/beit3 ← CLIP()/VLMo()/BEiT-3()
clip/vlmo/beit3.load_pretrained_weights()
Step 3: Generate embeddings with CLIP/VLMo/BEiT-3 encoder
image_embeddings, text_embeddings, image_text_embeddings
    ← clip/vlmo/beit3.generate_embeddings(data_loader)
Step 4: Integrate embeddings into the recommendation model
rec_model ← REC(image_embeddings, text_embeddings, image_text_embeddings)
Step 5: End-to-end training
for epoch in range(num_epochs) do
    # Forward pass
    user_item_scores ← rec_model.forward()
    # Compute loss
    loss ← compute_loss(user_item__scores, ground_truth)
    # Backward pass and optimisation
    optimiser.zero_grad()
    loss.backward()
    optimiser.step()
    # Update model with new embeddings
    rec_model.update(image_embeddings, text_embeddings, image_text_embeddings)
    # Evaluate and print performance metrics
    evaluate_and_print_metrics(epoch, rec_model)
end for
```

- RQ 2: How do parameter-efficient methods improve the efficiency of multimodal learning frameworks? (Chapter 6)

- RQ 3: How does our proposed multimodal learning framework perform in real-world scenarios? (From Chapter 7 to Chapter 9)

By addressing the three research questions, we validate the core thesis statement that enhancing intra-modal and inter-modal alignment in multimodal learning frameworks significantly improves performance across various tasks. RQ 1 evaluates how effectively our proposed contrastive learning methods tackle the issues of shallow intra-modal and inter-modal alignment within our MCA framework. Solving these alignment issues is expected to yield notable performance improvements. Additionally, as the size of neural models grows, the costs associated with tuning and running these models become prohibitively high, limiting the practical application of multimodal frameworks. Therefore, RQ 2 focuses on assessing how our parameter-efficient methods, such as the multi-way adapter, can mitigate efficiency issues and broaden the applicability of the MCA framework. Finally, we move beyond standard evaluations to apply and validate our framework in real-world scenarios. Through comprehensive evaluations in Chapters 7 to 9, we demonstrate the practical effectiveness of our framework in four critical applications, thereby addressing RQ 3. This systematic approach confirms that our enhancements to multimodal learning not only improve performance but also ensure scalability and real-world applicability, validating the overarching thesis statement.

4.4 Linkage Between Use-Case Applications and the Proposed Multimodal Learning Framework

The proposed multimodal learning framework is designed to address challenges inherent in processing and integrating information from diverse modalities, such as text, images, audio, and sensory data. Its relevance is demonstrated through various use-case applications, where the framework's ability to create semantically aligned and robust representations across modalities enhances task performance. This section clarifies the connections between these applications and the framework's design principles.

4.4.1 Addressing Domain-Specific Challenges

Each use-case application introduces unique domain-specific challenges, which are directly addressed by the framework:

Cross-Modal Retrieval: Applications such as image-text retrieval present significant challenges due to the inherent differences between visual and textual data. Images are represented in pixel-based spatial formats, while text is inherently sequential and symbolic, making direct comparisons between the two modalities non-trivial. Additionally, variations in linguistic descriptions, ambiguity in textual representations, and the contextual nature of images further complicate the alignment process. The proposed framework addresses these challenges by leveraging contrastive learning, which maps representations from different modalities into a shared semantic space. This shared space ensures that semantically similar image-text pairs are closely aligned while maintaining clear separation from dissimilar pairs. By optimizing this alignment, the framework improves retrieval efficiency, as evidenced by metrics such as Precision@K, which reflect better matching between images and their corresponding textual descriptions.

Multimodal Reasoning: In scenarios like crisis response, multimodal reasoning tasks require the integration of diverse inputs, such as textual reports and visual imagery, to derive actionable insights. These tasks are particularly challenging because the inputs often come from heterogeneous sources, may contain missing or noisy data, and require contextual understanding across modalities. Textual data might describe an event in detail, while visual data provides spatial and situational cues that are complementary but not explicitly linked. The framework tackles these challenges by incorporating attention-based mechanisms that selectively weigh and combine modality-specific features. This fusion not only ensures that relevant information is prioritized but also facilitates coherent reasoning by aligning the contextual relationships between text and images. Such mechanisms enable the generation of accurate and reliable predictions, critical for decision-making in high-stakes environments.

Robotic Vision: Robotic systems operating in real-world environments must process and integrate multimodal sensor data, such as visual inputs from cameras and tactile feedback from sensors. Challenges arise due to the differing nature and temporal characteristics of these modalities. For instance, visual data provides detailed spatial information but lacks tactile context, while touch data offers insights into texture or force but lacks spatial coverage. Furthermore, synchronizing these inputs and interpreting them in a unified manner is non-trivial, especially in dynamic or unstructured environments. The framework addresses these issues by encoding sensory inputs into a unified representation that captures both modality-specific and shared features. This integrated representation enhances the robot's ability to perceive complex scenes, infer context, and respond effectively. By improving multimodal integration, the framework significantly enhances task performance, such as object manipulation, navigation, or interaction with uncertain and variable surroundings.

4.4.2 Key Components and Their Applications

The core components of the framework are intrinsically linked to specific use cases:

- **Shallow Alignment:** Applicable in domains where lightweight models are required (e.g., edge devices in IoT systems), shallow alignment ensures computational efficiency while achieving sufficient modality integration.

- **Deep Alignment:** Suitable for high-complexity tasks like multimodal content generation or medical image-text analysis, deep alignment creates rich representations that capture intricate cross-modal relationships.

- **Contrastive Learning:** The framework's contrastive learning component aligns paired multimodal data, enabling effective performance in zero-shot and few-shot tasks such as cross-modal classification or retrieval in resource-constrained environments.

4.4.3 Generalization Across Use Cases

A key strength of the framework lies in its capacity to generalize across a wide range of applications, achieved through its flexible and adaptive design. The modular architecture of the framework plays a central role in this adaptability. Specific components, such as alignment mechanisms or fusion strategies, can be customized to suit the requirements of different modalities and domains. This modularity ensures that the framework remains versatile and scalable for various tasks. Furthermore, the creation of a shared representation space enables effective transfer learning. By aligning diverse inputs in a unified semantic space, the framework facilitates seamless adaptation to new tasks with minimal retraining. This ability to generalize makes it particularly well-suited for applications involving heterogeneous and evolving datasets.

4.4.4 Empirical Evidence of Linkage

The effectiveness of the framework is validated through empirical results on benchmark datasets, demonstrating its robustness and performance across diverse use cases. In cross-modal retrieval tasks, the framework achieves state-of-the-art Recall@K scores, reflecting its proficiency in aligning semantically similar data across modalities. For crisis response applications, it outperforms existing baselines with higher accuracy and F1 scores, showcasing its capability to integrate and reason over diverse modalities in high-stakes scenarios. Additionally, in robotic vision contexts, simulations reveal significant improvements in task performance, attributed to the benefits of enhanced multimodal pretraining. These results collectively highlight the framework's ability to address the unique challenges of different domains while maintaining high levels of efficiency and accuracy.

4.4.5 Conclusion

The proposed multimodal learning framework provides a versatile and robust solution for various use-case applications. By addressing domain-specific challenges, leveraging key components, and generalizing effectively across tasks, it establishes a strong linkage between theoretical advancements in multimodal learning and practical applications. This linkage underscores the framework's potential for broader adoption and impact in real-world scenarios.

Chapter 5

Improving Vision Performance through Contrastive Learning

5.1 Introduction

In Chapter 1, we discussed the importance of multimodal models in encoding visual inputs to generate high-quality visual embeddings, which are crucial for superior performance in both vision-specific and multimodal tasks. However, as stated in the thesis statement (Section 1.2), shallow intra-modal and inter-modal alignments limit the improvement in embedding quality and impact the performance in multimodal downstream tasks, such as Visual Question Answering (VQA) and Natural Language Video Retrieval (NVLR). Previous studies have explored contrastive learning methods to deepen intra-modal and inter-modal alignment, improving the quality of embeddings and boosting performance in both unimodal and multimodal contexts. Consequently, our thesis proposes several contrastive learning techniques designed to enhance intra-modal alignment, setting the stage for later applications in different domains. This approach aims to improve the model's generalization across various data types and its overall performance.

As outlined in Section 2.4, developing a robust image encoder, such as Vision Transformers (ViTs), is essential for advancing multimodal frameworks. Vision Transformers (ViTs) have emerged as popular models in computer vision, demonstrating state-of-the-art performance across various tasks, such as object identification and segmentation, highlighted by the seminal ViT model [224]. This success typically follows a two-stage strategy involving pre-training on large-scale datasets using self-supervised signals, such as masked random patches, followed by fine-tuning on task-specific labeled datasets with cross-entropy loss. Despite their state-of-the-art performance in various vision tasks, some studies [49, 292] indicate that this reliance on cross-entropy loss has been identified as a limiting factor in ViTs, affecting their generalization and transferability to downstream tasks.

Additionally, we chose not to include a text-only chapter because the out-of-the-box performance of state-of-the-art text encoders, such as those based on transformer models like BERT and GPT, is already very high. These models demonstrate exceptional performance across a wide range of natural language processing tasks, reducing the necessity for further fine-tuning specifically for text-only applications within the scope of this thesis.

Addressing this challenge, we introduce a novel Label-aware Contrastive Training framework, *LaCViT* in Section 5.2, which significantly enhances the quality of embeddings in ViTs. *LaCViT* not only addresses the limitations of cross-entropy loss but also facilitates more effective transfer learning across diverse image classification tasks. Our comprehensive experiments on eight standard image classification datasets reveal that *LaCViT* statistically significantly enhances the performance of three evaluated ViTs by up-to 10.78% under Top-1 Accuracy.

Moreover, while recent works employing contrastive learning address some of these limitations by enhancing the quality of embeddings and producing better decision boundaries, they often overlook the importance of hard negative mining and rely on resource intensive and slow training using large sample batches. Indeed, integrating contrastive learning at the fine-tuning stage introduces a dependence on large mini-batch sizes (e.g., 1024 or 2048), unlike methods based on cross-entropy. To counter these issues, we introduce a novel approach named CLCE as detailed in Section 5.3, which integrates Label-Aware Contrastive Learning with CE. Our approach not only maintains the strengths of both loss functions but also leverages hard negative mining in a synergistic way to enhance performance. Experimental results demonstrate that CLCE significantly outperforms CE in Top-1 accuracy across twelve benchmarks, achieving gains of up to 3.52% in few-shot learning scenarios and 3.41% in transfer learning settings with the BEiT-3 model. Importantly, our proposed CLCE approach effectively mitigates the dependency of contrastive learning on large batch sizes such as 4096 samples per batch, a limitation that has previously constrained the application of contrastive learning in budget-limited hardware environments.

Lastly, human-annotated vision datasets inevitably contain a fraction of human-mislabelled examples, often due to human error when one class superficially resembles another. While the detrimental effects of such mislabelling on supervised learning are well-researched, their influence on Supervised Contrastive Learning (SCL) remains largely unexplored. The efficacy of Supervised Contrastive Learning (SCL) hinges on the quality of supervision labels. Labelling errors can lead to incorrect positive and negative pairings, undermining the integrity of the representations learned. In Section 5.4.4, we show that human-labelling errors not only differ significantly from synthetic label errors, but also pose unique challenges in SCL, different to those in traditional supervised learning methods. Specifically, our results indicate they adversely impact the learning process in the $\sim$99% of cases when they occur as false positive samples. Existing noise-mitigating methods primarily focus on synthetic label errors and tackle the unrealistic setting of very high synthetic noise rates (40–80%), but they often underperform on common image datasets due to overfitting. To address this issue, we introduce a novel SCL objective with robustness to human-labelling errors, SCL-RHE in Section 5.4, facilitating better use of existing datasets without the need for labor-intensive manual re-annotation. SCL-RHE is designed to mitigate the effects of real-world mislabelled examples, typically characterized by much lower noise rates ($< 5\%$). We demonstrate that SCL-RHE consistently outperforms state-of-the-art representation learning and noise-mitigating methods across various vision benchmarks, by offering improved resilience against human-labelling errors.

Relevance of Optimizing Contrastive Learning in Multimodal Learning

Contrastive learning is a pivotal optimization strategy in multimodal learning, primarily due to its capacity to align and integrate representations from diverse modalities. It encourages cross-modal coherence by minimizing the distance between representations of semantically similar (positive) pairs while maximizing the distance between dissimilar (negative) pairs. This process creates a shared representation space, ensuring semantic alignment across modalities and enabling effective cross-modal interactions. Furthermore, it addresses modality-specific disparities that arise from distinct feature distributions and encoding mechanisms in modalities like text, images, and audio. By optimizing for semantic alignment, contrastive learning mitigates these disparities, ensuring related inputs are closely aligned in the representation space regardless of their modality.

Another critical advantage of contrastive learning is its ability to enhance robustness to noise and missing data. Multimodal contexts often involve scenarios where one modality may be noisy or entirely absent. Contrastive learning helps models prioritize meaningful cross-modal relationships over modality-specific noise, resulting in robust and generalizable representations. This robustness directly translates to improved performance in cross-modal tasks. For instance, in cross-modal retrieval, semantically aligned representations enable effective retrieval of items across modalities, such as matching text descriptions with relevant images. Similarly, visual-language models, including CLIP and ALIGN, leverage contrastive learning to achieve state-of-the-art results in tasks such as zero-shot image classification and multimodal reasoning.

Additionally, contrastive learning facilitates few-shot and zero-shot learning by exploiting inherent relationships between modalities to build shared representations that generalize effectively to unseen data. This makes it particularly valuable for multimodal scenarios with limited training examples. The scalability and efficiency of contrastive learning further underscore its relevance. By generating positive and negative pairs automatically from natural correspondences, such as image-text pairs in datasets, it eliminates the need for extensive manual labeling, enabling efficient training of large-scale multimodal models.

In conclusion, optimizing contrastive learning effectively aligns multimodal data, enhances robustness, and improves the performance of multimodal systems across diverse tasks. Its scalable and efficient design has made it a foundational approach in multimodal machine learning frameworks.

5.2 LaCViT: A Label-aware Contrastive Training Framework for Vision Transformers

5.2.1 Introduction

Transformers have significantly advanced the field of computer vision, particularly in tasks such as image classification [115, 143, 146, 147, 224, 242, 271]. These models typically follow a two-stage process: pre-training on auxiliary tasks and fine-tuning on specific tasks using cross-entropy loss. However, the reliance on cross-entropy often leads to poor generalization and vulnerability to label noise and adversarial attacks [10, 22, 136, 144, 145, 166], which impede their efficacy in practical applications. Moreover, vision transformers exhibit a lack of learned inductive biases [49, 149], an essential feature for handling unseen examples and enhancing transfer learning.

The inherent lack of inductive bias and the limitations of fine-tuning with cross-entropy compromise the transfer learning capabilities of vision transformers, particularly when the target domain has a small sample size [292]. Although previous works have attempted to address these issues by integrating convolutional neural networks or modifying the transformer architecture [71, 238, 263], these solutions often compromise the inherent advantages of transformers, like their training efficiency and scalability. Hence, it would be advantageous to have an alternative approach to improve the transfer effectiveness of vision transformers without relying on convolutional models or layers, while utilizing task labels in the fine-tuning stage.

The development of contrastive learning trace back to early explorations by Becker [16]. This approach aims to differentiate similar items from dissimilar ones within an embedding space. Additionally, contrastive learning has shown remarkable efficacy in improving deep learning model performance across various domains [142, 143], including sentence [67, 134] and audio representation learning [294], with its most notable impact observed in image recognition tasks, as exemplified by SimCLR [29] and other studies [144, 145, 149]. While integrating label information into contrastive learning has been explored, as in [109], these efforts have primarily remained confined to the pre-training phase and have not been extended to vision transformers.

While both fields have advanced in parallel, the integration of label-aware contrastive learning within the fine-tuning stage of vision transformers remains unexplored. Our work addresses this gap by pioneering the application of contrastive learning during the fine-tuning phase of vision transformers, thereby enhancing their transferability.

In response, we propose the *LaCViT*, a label-aware contrastive training framework designed specifically for vision transformers. *LaCViT* leverages task labels in a contrastive learning context to fine-tune pre-trained models, thus significantly improving their transfer learning capabilities. This framework employs a two-stage, label-aware contrastive learning loss to refine sample embeddings, enabling better generalization to target tasks. Notably, *LaCViT* is the first framework of its kind to enhance vision transformer transfer learning without relying on convolutional layers or extended training epochs. We believe that our work serves as an impetus for the research community to reconsider the fine-tuning mechanisms for Vision Transformers. The primary contributions of this section are as follows:

- We introduce *LaCViT*, a pioneering label-aware contrastive fine-tuning framework that substantially enhances the transfer learning capabilities of vision transformers, addressing the thesis statement's focus on improving intra-modal and inter-modal alignment.
- We demonstrate the wide applicability of *LaCViT* by fine-tuning three vision transformer models, validating its versatility and effectiveness across different configurations.
- Extensive experimentation across eight image classification datasets shows that *LaCViT* significantly outperforms baseline models. For example, it achieves a 10.78% increase in Top-1 Accuracy for the *LaCViT*-trained MAE on the CUB-200-2011 dataset [80], supporting our thesis claim of enhanced performance through improved alignment.
- Additional analysis reveals that *LaCViT* effectively reshapes pretrained embeddings into a more discriminative space, further enhancing performance on target tasks and corroborating our thesis statement on the importance of deep alignment.
- The original material in this section has been accepted for presentation at the 2024 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), a conference with an h5-index of 123.

5.2.2 The Proposed Approach

Motivation and Overview

The prevalent approach for fine-tuning vision transformers employs cross-entropy loss, which suffers from poor generalization capabilities. Moreover, existing contrastive learning methods overlook the utility of label information in the fine-tuning phase. To address these limitations, we propose *LaCViT*, a label-aware contrastive fine-tuning framework designed to enhance the transfer learning capabilities of vision transformers. As illustrated in Figure 5.1, our *LaCViT* consists of two distinct stages: *the label-aware contrastive training stage* and *the task head fine-tuning stage*.

- **Label-aware Contrastive Training Stage (Stage 1)**: In this stage, we initialize the model with pretrained weights and adapt these weights through a contrastive learning loss that incorporates the label information of the target task. This process comprises four main

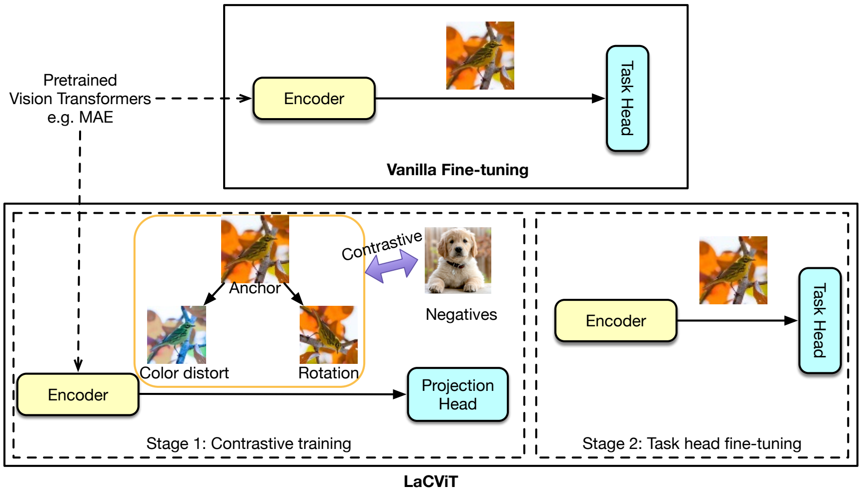


Figure 5.1: **The overview of *LaCViT*, which consists of two training stages: 1) label-aware contrastive training and 2) task head fine-tuning**, compared to the vanilla fine-tuning, which directly fine-tunes the task head. The first contrastive training stage trains the vision transformers based on the labels of the target tasks with a contrastive loss, aiming to improve the embedding quality, and in the second stage, *LaCViT* is fine-tuned with a task-specific head.

steps: data augmentation, patch encoding, nonlinear projection, and contrastive loss computation.

- **Task Head Fine-tuning Stage (Stage 2)**: The second stage focuses on training the task-specific head, typically a simple linear layer for classification tasks. This stage utilizes a standard cross-entropy loss to fine-tune the whole framework, which is added atop the pretrained vision transformer.

Label-aware Contrastive Training Stage

Data Augmentation: To augment each image in the mini-batch into two transformed views, we employ AutoAugment [41], which has proven to be highly effective for contrastive learning.

Encoding: Feature embeddings for each of the two augmented views of the image are generated using an encoder, such as ViT [224], MAE [115], or SimMIM [262].

Nonlinear Projection Head: To enhance the quality of the embeddings, we employ a nonlinear projection head, $g(h)$, upon the encoder to map the representation to the space where the contrastive loss is applied. Thus, to implement $z_i = g(h_i) = W^{(2)}\sigma(W^{(1)}h_i)$, we use two dense layers, where σ is a ReLU function. The $z = g(h)$ is trained to be invariant to data transformation, which means g removes information that could be useful for the downstream task (e.g., color of objects). By using the nonlinear projection, more information can be maintained in h. These embeddings are then grouped into distinct sets by the training class label of the source image.

Contrastive Loss Objective: To obtain a more discriminative representation space for the target

			CIFAR-10	CIFAR-100	Cub-200-2011	Oxford-Flowers	Oxford-Pets	iNat 2017	ImageNet-1k	Places365
Model	**Seen dataset**	**FT method**	Acc@1	Acc@1	Acc@1	Acc@1	Acc@1	Acc@1	Acc@1	Acc@1
Data2vec	ImageNet-21k	CE	98.25	89.21	85.16	91.57	94.52	71.05	**84.20**	58.73
ViT-B	ImageNet-1k	CE	98.13	87.13	N/A	89.49	93.81	65.26	77.91	54.06
LaCViT-ViT-B	ImageNet-21k	*LaCViT*	99.07	91.01	85.69	**94.98**	94.57	70.38	82.99	57.73
ViT-L	ImageNet-1k	CE	97.86	86.36	N/A	89.66	93.64	64.82	76.53	54.55
SimMIM	ImageNet-1k	CE	98.78	90.26	76.47	83.46	94.22	70.28	83.00	57.54
LaCViT-SimMIM	ImageNet-1k	*LaCViT*	99.11	90.80	85.79	92.25	94.85	71.34	83.64	58.47
MAE	ImageNet-1k	CE	98.28	87.67	78.46	91.67	94.05	70.50	83.60	57.90
MAE	ImageNet-1k	SimCLR	97.53	76.01	57.91	89.22	91.15	65.62	81.92	55.48
MAE	ImageNet-1k	N-pair-loss	95.23	73.76	52.56	89.87	87.12	62.36	78.34	52.97
LaCViT-MAE	ImageNet-1k	*LaCViT*	**99.34**	**91.27**	**89.24**	93.34	**95.63**	**72.55**	84.12	**58.92**

Table 5.1: **Image classification performance benchmarks over eight datasets.** CE refers to fine-tune with cross-entropy, while *LaCViT* refers to fine-tune with our proposed label-aware contrastive fine-tuning framework.

task, we train the pretrained encoder using a contrastive loss by leveraging label information (i.e., the label-aware contrastive loss). The label-aware contrastive loss enables stronger geographic clustering of samples belonging to the same class in the embedding space, while simultaneously pushing apart clusters of samples from different classes. The advantage of the label-aware contrastive loss is that we compute the contrastive loss based on true positive pairs per anchor in addition to true negative samples, compared to self-supervised contrastive learning that uses only augmented views. The contrastive loss is mathematically defined as follows:

$$\mathscr{L}(\mathscr{D}^*) = \sum_{z_i \in \mathscr{D}^*} \frac{-1}{|\mathscr{D}^{*+}_{-z_i}|} \sum_{z_p \in \mathscr{D}^{*+}_{-z_i}} \log \frac{\exp(z_i \cdot z_p/\tau)}{\sum\limits_{z_a \in \mathscr{D}^{*}_{-z_i}} \exp(z_i \cdot z_a/\tau)}, \tag{5.1}$$

In Equation 5.4, $\mathscr{D}^*$ represents the entire mini-batch composed of an embedding z for each image view (or anchor) i. Therefore, $z_i \in \mathscr{D}^*$ is a set of embeddings within the mini-batch. The superscripts $+$ and $-$, e.g. $\mathscr{D}^{*+}$, denote sets of embeddings consisting only of positive and negative examples, respectively, for the current anchor within the mini-batch. The term $|\mathscr{D}^{*+}_{-z_i}|$ represents the cardinality of the positive set for the current anchor, while the subscript $-z_i$ denotes that this set excludes the embedding z_i. The symbol $\cdot$ represents the dot product. τ is a temperature parameter, which controls the degree of loss applied when two images have the same class but the embeddings are different. A higher value pushes the model to more strongly separate the positive and negative examples.

Comparison with Existing Methods. Different from previous methods that integrate convolutional layers into the transformer architecture or extend the training epochs, our *LaCViT* preserves the native advantages of transformers such as training efficiency and enhances their transfer learning capabilities through label-aware contrastive training.

5.2.3 Experiments

Experimental Setup: To evaluate the effectiveness of the *LaCViT* framework, we conducted experiments using three state-of-the-art pretrained vision transformer models across eight diverse image classification datasets. The train/test splits for these datasets are consistent with prior work [49]. For the contrastive training stage, we initialize the encoder with pretrained weights obtained from either the ImageNet-1k or ImageNet-21k datasets. During training, we employ a batch size of 4096 for the contrastive training stage (Stage 1) and 128 for the task head fine-tuning stage (Stage 2). The number of epochs for these stages is set to 50 and 10, respectively. Both stages use an initial learning rate of 1×10^{-4}. The temperature parameter τ for the contrastive loss is set to 0.1. All the code used in our experiments can be found in `https://github.com/longkukuhi/LaCViT`.

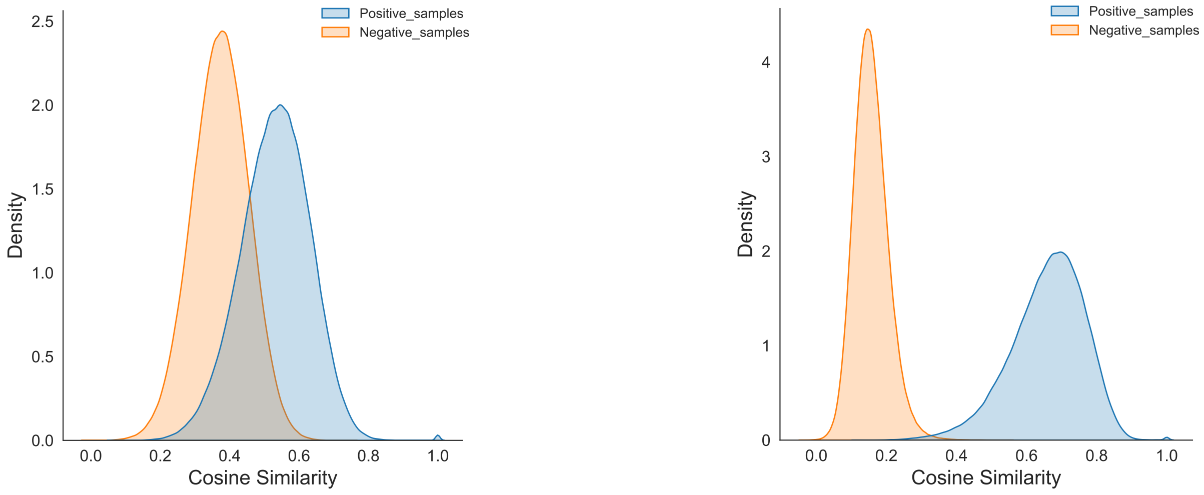
Comparative Analysis: We benchmark the performance of *LaCViT*-trained models against baseline models that solely utilize vanilla cross-entropy loss. The evaluation metrics include Top-1 accuracy across the selected datasets, as summarized in Table 5.1. Our results indicate that *LaCViT*-trained models consistently outperform their baseline counterparts. For instance, *LaCViT*-ViT-B achieves a Top-1 accuracy improvement of 3.88% and 5.49% on the CIFAR-100 and Oxford 102 Flowers datasets, respectively. *LaCViT*-SimMIM shows a significant advantage over SimMIM, with an average improvement of 2.74% over tested datasets. Data2vec performs worse than *LaCViT*-MAE on smaller datasets such as CUB-200-2011 but shows marginal improvement on larger datasets, attributable to its larger pre-training dataset (ImageNet-21k). Moreover, *LaCViT*-MAE emerges as the best-performing model on almost all datasets, with a performance boost of 10.78% on the CUB-200-2011 dataset[1].

Discussion: The observed performance gains substantiate the effectiveness of our label-aware contrastive training approach in *LaCViT*. This is particularly prominent in smaller datasets but is also evident in larger datasets, such as iNat 2017 and ImageNet-1k. *LaCViT*-MAE achieves better performance with less pre-training data, which demonstrates that extensive pre-training on larger datasets does not necessarily translate to improved transferability on smaller datasets. With *LaCViT*, comparable or even superior performance can be attained.

5.2.4 Analysis

Ablation Study on Alternative Contrastive Loss: We evaluate the performance of *LaCViT* against other notable unsupervised contrastive learning methods, namely SimCLR [29] and N-pair-loss [217], to understand the unique advantages of our label-aware approach. We use the MAE base model for these comparisons. The results are summarized in the lower section of Table 5.1. MAE fine-tuned with SimCLR significantly underperforms compared to *LaCViT*-MAE, particularly on the CUB-200-2011 and CIFAR-100 datasets, registering a performance decrease up to 31.33%. When fine-tuned with N-pair-loss, MAE exhibits a 2–5% decline in accuracy compared to its SimCLR counterpart, with the exception of a marginal accuracy gain of 0.65% on the Oxford 102 Flower dataset. These findings suggest that unsupervised contrastive learning methods may not sufficiently capture class-specific features, thus affecting performance

[1] An exception is the Oxford-Flowers dataset, where *LaCViT*-ViT-B excels due to its ImageNet-21k pre-training.



(a) MAE

(b) *LaCViT*-MAE

Figure 5.2: **Plot of cosine similarity distribution across two random classes from CIFAR-10.** Blue and orange mean positive and negative similarities, respectively.

adversely. Cross-entropy fine-tuned MAE consistently outperforms both SimCLR and N-pair-loss fine-tuned versions, emphasizing the need for using label information in contrastive learning like *LaCViT*.

Embedding Quality Analysis: We further investigate the geometric properties of the learned embedding spaces to understand the impact of label-aware contrastive training.

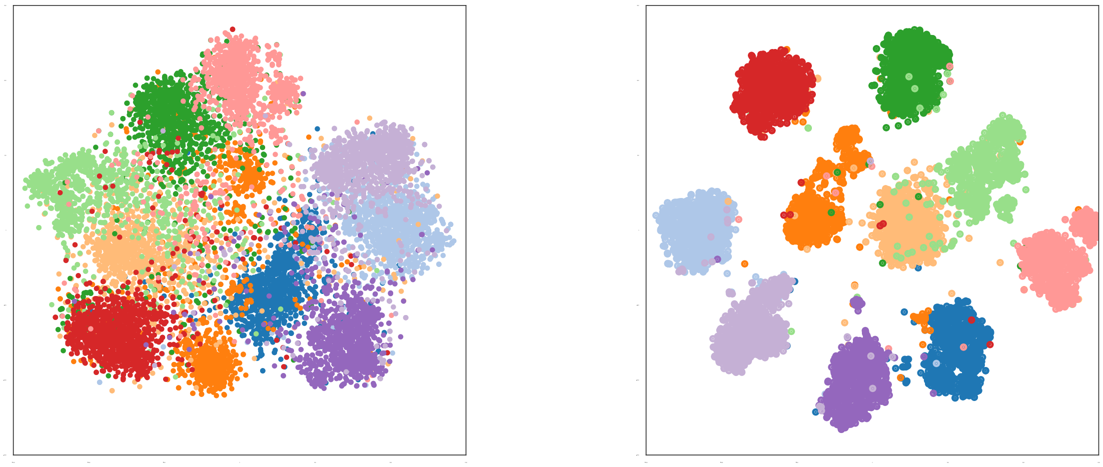
- **Cosine Similarity**: Figure 5.2 illustrates the cosine similarity distribution between *LaCViT*-MAE and MAE. The figure shows that *LaCViT*-MAE offers better inter-class separation.

- **t-SNE Visualization**: Figure 5.3 presents t-SNE visualizations of the embeddings for both MAE and *LaCViT*-MAE. The clusters in *LaCViT*-MAE are tighter and better separated, thus underscoring the discriminative power of label-aware contrastive training.

Summary: Our analysis confirms that label-aware contrastive training with *LaCViT* enhances the geometric properties of the embedding spaces, particularly in terms of inter-class separation. These improvements substantiate the superior transfer learning capabilities of vision transformers trained using *LaCViT*.

5.2.5 Summary

In this section, we present *LaCViT*, a label-aware contrastive fine-tuning framework that significantly increases the Top-1 accuracy of vision transformers across various benchmarks. It outperforms state-of-the-art models like MAE by up to 10.78% and is applicable to other transformers such as ViT and SimMIM. Through rigorous analysis, including using cosine similarity



(a) MAE

(b) *LaCViT*-MAE

Figure 5.3: **Embedding Space Visualization for MAE vs. *LaCViT*-MAE**. Displayed over ten CIFAR-10 classes using t-SNE. Each dot represents a sample, with distinct colors indicating different label classes.

metrics and t-SNE visualizations, we demonstrate that *LaCViT* effectively reshapes the geometric properties of the embedding space, contributing to its effectiveness in image classification tasks. In summary, *LaCViT* offers a comprehensive and versatile approach that serves to substantially elevate the utility of transformers in image classification. Our exhaustive empirical evaluations not only validate the effectiveness of *LaCViT* but also suggest that it offers an effective alternative to cross-entropy for fine-tuning pretrained models for image classification tasks, addressing the thesis statement's focus on improving intra-modal and inter-modal alignment.

5.3 CLCE: An Approach to Refining Cross-Entropy and Contrastive Learning for Optimized Learning Fusion

In Section 5.2, we validated our initial method, *LaCViT*, which enhances transferability via contrastive learning during the fine-tuning phase. The promising outcomes of this approach have motivated further exploration into optimizing contrastive learning to achieve superior performance.

Although we achieved promising results of using our *LaCViT*, the integrating contrastive learning at the fine-tuning stage introduces a dependence on large mini-batch sizes (e.g., 1024 or 2048), unlike methods based on cross-entropy.

As discussed in Section 5.2.1, approaches for achieving state-of-the-art performance in image classification tasks often employ models initially pre-trained on auxiliary tasks and then fine-tuned on a task-specific labeled dataset with a Cross-Entropy loss (CE) [115, 224, 252].

However, CE's inherent limitations can impact model performance. Specifically, the measure of KL-divergence between one-hot label vectors and model outputs can cause narrow decision margins in the feature space. This hinders generalization [22, 136] and has been shown to be sensitive to noisy labels [136, 166] or adversarial samples [56, 166]. Various techniques have emerged to address these problems, such as knowledge distillation [83], self-training [264], Mixup [282], CutMix [277], and label smoothing [229]. However, in scenarios such as few-shot learning, these issues with CE have not been fully mitigated. Indeed, while techniques such as extended fine-tuning epochs and specialized optimizers [162, 284] can reduce the impact of CE to some extent, they introduce new challenges, such as extended training time and increased model complexity [162, 284].

Amidst these challenges in context of image classification, contrastive learning has emerged as a promising solution, particularly in few-shot learning scenarios such as CIFAR-FS [18] and CUB-200-2011 datasets [245]. The effectiveness of contrastive learning lies in its ability to amplify similarities among positive pairs (inter-class data points) and distinguish negative pairs (inter-class data points). SimCLR [186], for instance, has utilized instance-level comparisons unsupervised. However, this unsupervised approach raises concerns regarding its effectiveness, primarily because it limits the positive pairs to be transformed views of an image and treats all other samples in a mini-batch as negatives, potentially overlooking actual positive pairs. We hypothesis incorporating task-specific label information is thus crucial for accurately identifying all positive pairs, especially given the presence of labels in many downstream datasets.

There is a growing trend of using task labels with contrastive learning to replace the standard use of CE [109]. A critical observation here is that many state-of-the-art methods, both in supervised [74, 109] and unsupervised [64, 186, 217] contrastive learning, overlook the strategic selection of negative samples. They fail to differentiate or prioritize these samples during selection or processing, thereby missing the benefits of leveraging "hard" negative samples, as highlighted in numerous studies [39, 79, 120, 171, 226, 287]. While contrastive learning mitigates the limitations of CE, it simultaneously introduces a challenge: a reliance on large batch sizes—such as 2048 or 4096 samples per batch—for superior performance compared to CE. This requirement is often impractical in budget hardware environments, particularly when using GPUs with less than 24 GB of memory. As a consequence, state-of-the-art methods such as SupCon [109] underperform compared to CE when using more commonly employed batch sizes, such as 64 or 128 samples per batch, which limits their application. Motivated by these successes and gaps in research, we pose the question: *How can the performance of contrastive learning be improved to address the shortcomings of cross-entropy loss, while also mitigating the reliance on large batch sizes?*

Building upon the identified research gaps, in this Section, we propose CLCE, an innovative approach that combines Label-Aware Contrastive Learning with CE. This approach effectively merges the strengths of both loss functions and integrates hard negative mining. This technique refines the selection of positive and negative samples, thereby enabling CLCE to achieve state-of-the-art performance. As our empirical findings illustrate in Fig. 5.4, CLCE places a greater emphasis on hard negative samples that are visually very similar to positive samples, forcing the encoder to learn how to generate more distinct embeddings and better decision boundaries. The core contributions of this Section can be summarised as follows:

- Introduction of an Innovative Approach:We introduce CLCE, a groundbreaking method that enhances model performance without requiring specialized architectures or additional resources. Our work is the first to successfully integrate explicit hard negative mining into Label-Aware Contrastive Learning, retaining the benefits of cross-entropy (CE) while eliminating the dependence on large batch sizes. This contribution directly addresses the thesis statement's focus on improving multimodal learning efficiency and alignment.

- State-of-the-Art Performance in Few-Shot and Transfer Learning Settings: CLCE significantly surpasses CE, achieving an average of 2.74% higher Top-1 accuracy across four few-shot learning datasets using the BEiT-3 base model [252], with notable gains in 1-shot learning scenarios. Additionally, in transfer learning settings, CLCE consistently outperforms other state-of-the-art methods across eight image datasets, establishing a new benchmark for base models (88 million parameters) on ImageNet-1k [45]. These results validate our thesis claim of enhanced performance through improved alignment.

- Reduced Dependency on Large Batch Sizes in Contrastive Learning: Empirical evidence shows that CLCE significantly outperforms both CE and previous state-of-the-art contrastive learning methods like SupCon [109] even with commonly used batch sizes, such as 64, where earlier methods underperform. This advancement addresses a critical bottleneck in contrastive learning, particularly in resource-limited settings, aligning with our thesis goal of creating more efficient and practical multimodal learning methods. CLCE emerges as a viable, efficient alternative to conventional CE, further supporting the thesis statement.

- The original material in this section has been accepted for presentation at the The 27th European Conference on Artificial Intelligence (Core ranking A Conference) as a full paper.

5.3.1 Approach

In this Section, we propose an enhanced approach named CLCE for image models that integrates our propose Label-Aware Contrastive Learning with the Hard Negative Mining (LACLN) and the Cross-Entropy (CE). CLCE harnesses the potential of contrastive learning to mitigate the limitations inherent in CE while preserving its advantages. Specifically, LACLN enhances similarities between instances of the same class (i.e. positive samples) using label information

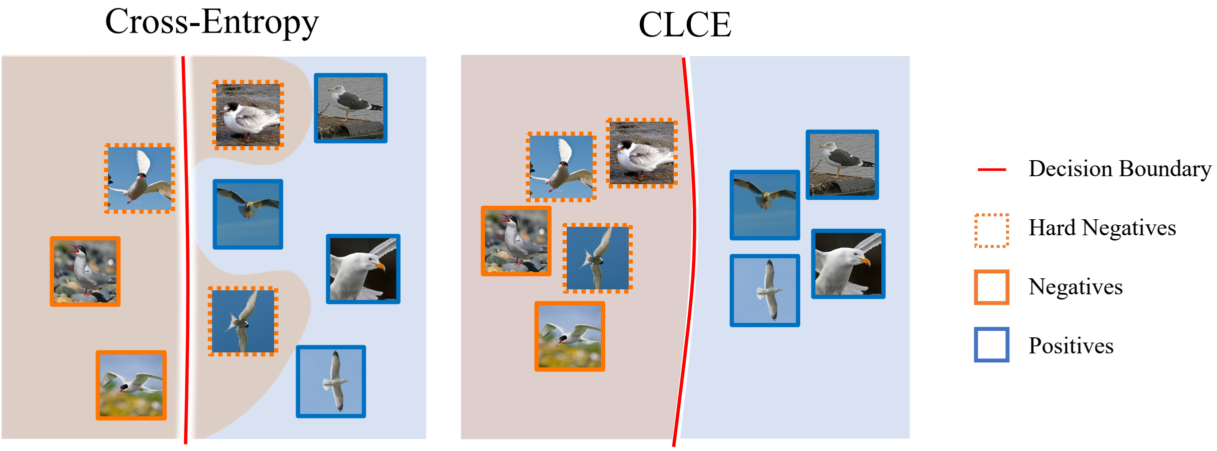


Figure 5.4: CLCE, our proposed approach, integrates a Label-Aware Contrastive Learning with the Hard Negative Mining (LACLN) term and a CE term. Illustrated with CUB-200-2011 dataset, it emphasizes hard negatives (thick dashed borders) for better class separation. This underscores their marked visual similarity to their positive counterparts. Blue indicates positive examples and orange denotes negatives. On the right, CLCE visibly separates class embeddings more effectively and results a better decision boundary than traditional CE.

and contrasts them against instances from other classes (i.e. negative samples), with particular emphasis on hard negative samples. Thus, LACLN reshapes pretrained embeddings into a more distinct and discriminative space, enhancing performance on target tasks. Moreover, CLCE's foundation draws from the premise that the training efficacy of negative samples varies between soft and hard samples. We argue that weighting negative samples based on their dissimilarity to positive samples is more effective than treating them equally. This allows the model to prioritize distinguishing between positive samples and those negative samples that the embedding deems similar to the positive ones, ultimately enhancing overall performance.

CLCE

The overall proposed CLCE approach is a weighted combination of LACLN and standard CE, as expressed in Eq. 5.2:

$$\mathscr{L}_{\text{CLCE}} = (1-\lambda)\mathscr{L}_{\text{CE}} + \lambda \mathscr{L}_{\text{LACLN}} \tag{5.2}$$

In Eq. 5.2, the term $\mathscr{L}_{\text{CE}}$ represents the CE loss, while $\mathscr{L}_{\text{LACLN}}$ symbolizes our proposed LACLN loss. λ represents a scalar weighting hyperparameter. λ determines the relative importance of each of the two losses. To provide context for $\mathscr{L}_{\text{CE}}$, we refer to the standard definition of the multi-class CE loss, detailed in Eq. 5.3:

$$\mathscr{L}_{\text{CE}} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} z_{i,c}\log(\hat{z}_{i,c}) \tag{5.3}$$

In Eq. 5.3, $z_{i,c}$ and $\hat{z}_{i,c}$ represent the label and the model's output probability for the ith instance belonging to class c, respectively.

$$\mathscr{L}_{\text{LACLN}} = \sum_{x_i \in \mathscr{D}^*} -\log \frac{1}{|\mathscr{D}^{*+}_{-x_i}|} \frac{\sum\limits_{x_p \in \mathscr{D}^{*+}_{-x_i}} \exp(x_i \cdot x_p/\tau)}{\sum\limits_{x_p \in \mathscr{D}^{*+}_{-x_i}} \exp(x_i \cdot x_p/\tau) + \sum\limits_{x_k \in \mathscr{D}^{*-}_{-x_i}} \frac{|\mathscr{D}^{*-}_{-x_i}|}{\sum\limits_{x_k \in \mathscr{D}^{*-}_{-x_i}} \exp(x_i \cdot x_k/\tau)} \exp(x_i \cdot x_k/\tau)^2} \tag{5.4}$$

We present the formal definition of our LACLN in Eq. 5.4. This loss introduces a weighting factor for each negative sample, calculated based on the dot product (indicating similarity) between the sample embeddings and the anchor, and normalized by a temperature parameter τ. This formulation strategically emphasizes “hard” negative samples — those closely associated with the positive samples by the model’s current embeddings. Specifically, the weighting factor for negative samples is determined by calculating their relative proportion based on the average similarity (dot product) observed within each mini-batch. The essence of Eq. 5.4 is to minimize the distance between positive pair embeddings and maximize the separation between the anchor and negative samples, particularly the hard negatives. This objective is achieved through two components: the numerator, focusing on bringing positive sample embeddings closer to the anchor, and the denominator, containing both positive and weighted negative samples to ensure the anchor’s embedding is distant from negative samples, with a special focus on the more challenging ones. The integration of hard negative mining into contrastive learning is critical as it sharpens the model’s ability to differentiate between closely related samples, thus enhancing feature extraction and overall model performance.

Specifically, $\mathscr{D}^*$ represents the entire mini-batch composed of an embedding x for each image view (or anchor) i. Therefore, $x_i \in \mathscr{D}^*$ is a set of embeddings within the mini-batch. The superscripts $+$ and $-$, e.g. $\mathscr{D}^{*+}$, denote sets of embeddings consisting only of positive and negative examples, respectively, for the current anchor within the mini-batch. The term $|\mathscr{D}^{*+}_{-x_i}|$ represents the cardinality of the positive set for the current anchor, while the subscript $-x_i$ denotes that this set excludes the embedding x_i. The symbol $\cdot$ represents the dot product. τ is a scalar temperature parameter controlling class separation. A lower value for τ encourages the model to differentiate positive and negative instances more distinctly.

The use of the square in $\exp(x_i \cdot x_k/\tau)$ is not ad hoc but follows from the structure of our weighting and scaling mechanism. The expression $\frac{\exp(x_i \cdot x_k/\tau)}{\sum_{x_k \in \mathscr{D}^{*-}_{-x_i}} \exp(x_i \cdot x_k/\tau)}$ serves as a weighting factor that normalizes the similarity score of each negative sample relative to the sum of all similarity scores in the batch. Then, we multiply this weight by the original similarity score $\exp(x_i \cdot x_k/\tau)$,which acts as the scaling. Thus, the operation can be represented as: $\frac{\exp(x_i \cdot x_k/\tau)}{\sum_{x_k \in \mathscr{D}^{*-}_{-x_i}} \exp(x_i \cdot x_k/\tau)} \cdot \exp\left(x_i \cdot x_k/\tau\right)$ This results in the squared term, which arises naturally from this weighting and scaling process. The weighting factor $\mathscr{D}^{*-}_{-x_i}$ is used to normalize the contribution of negative samples in a multi-viewed batch. This normalization serves to remove bias that may be present among the negatives, ensuring a balanced contribution to the loss. This approach is a proven technique used in many contrastive loss functions to enhance performance and stability.

Analysis of CLCE

Notably, our proposed CLCE has the following desirable properties:

- Robust Positive/Negative Differentiation: We ensure a clear distinction between true positive and true negative samples by leveraging explicit label information, as encapsulated in Eq. 5.4. This not only prevents the model from being misled by incorrectly contrasting of samples but also reinforces the core philosophy of contrastive learning. The aim is two-fold: to reduce the distance between the embeddings of positive pairs and to increase the distance for negative pairs, ensuring robust class separation.

- Discriminating Fine Detail with Hard Negatives: Our loss adjusts the weighting of negative samples based on their similarities to positive instances, as defined in Eq. 5.4. This nuanced approach ensures that the model not only differentiates between glaringly distinct samples but also adeptly distinguishes more challenging, closely related negative samples. Such an approach paves the way for a robust model that discerns real-world scenarios where differences between classes might be minimal.

Representation Learning Framework

We use a representation learning framework comprised of three main components, designed specifically to optimize our CLCE approach:

- **Data Augmentation module,** $Aug(\cdot)$: This component creates two different views of each sample x, denoted $\tilde{x} = Aug(x)$. This means that every sample will have at least one similar sample (positive pair) in a batch during training.

- **Encoder Network,** $Enc(\cdot)$: This network encodes the input data, x, into a representation vector, $r = Enc(x)$. Each of the two different views of the data is fed into the encoder separately.

- **Classification head,** $Head(\cdot)$: This maps the representation vector, r, to probabilities of classes in the target task. The mapping primarily consists of a linear layer, and we utilize its output to calculate the cross-entropy loss.

Our CLCE approach (Eq. 5.4) can be applied using a wide range of encoders, such as BEiT-3 [252] or the ResNets [193] for image classification. Following the method in [30], every image in a batch is altered to produce two separate views (anchors). Views with the same

		CIFAR-FS		FC100		miniImageNet		tieredImageNet	
Model	Loss	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot
[47]	Transductive	76.58±0.68	85.79±0.50	43.16±0.59	57.57±0.55	65.73±0.68	78.40±0.52	73.34±0.71	85.50±0.50
[285]	Meta-QDA	75.83±0.88	88.79±0.75	-	-	67.83±0.64	84.28±0.69	74.33±0.65	89.56±0.79
[82]	FewTRUE-ViT	76.10±0.88	86.14±0.64	46.20±0.79	63.14±0.73	68.02±0.88	84.51±0.53	72.96±0.92	87.79±0.67
[82]	FewTRUE-Swin	77.76±0.81	88.90±0.59	47.68±0.78	63.81±0.75	72.40±0.78	86.38±0.49	76.32±0.87	89.96±0.55
[91]	BAVARDAGE	82.68±0.25	89.97±0.18	52.60±0.32	65.35±0.25	77.85±0.28	88.02±0.14	79.38±0.29	88.04±0.18
ResNet-101	CE	69.80±0.84	85.20±0.62	43.71 ±0.73	58.65±0.74	55.73±0.85	73.86±0.65	46.93±0.85	62.93±0.76
ResNet-101	H-SCL [103]	67.25±0.86	84.51±0.65	41.34±0.72	57.02±0.70	53.38±0.79	70.29±0.63	44.43±0.82	60.83±0.71
ResNet-101	CE+SupCon	73.61±0.80	86.15±0.53	45.30±0.62	60.18±0.72	57.49±0.82	75.63±0.61	49.44±0.79	66.47±0.60
ResNet-101	CLCE (this work)	76.14±0.75	87.93±0.48	49.48±0.57	64.31±0.70	66.20±0.74	83.41±0.55	63.61±0.72	79.83±0.51
BEiT-3	CE	83.68±0.80	93.01±0.38	66.35±0.95	84.33±0.54	90.62±0.60	95.77±0.28	84.84±0.70	94.81±0.34
BEiT-3	H-SCL [103]	82.21±0.80	91.49±0.37	65.27±0.98	82.61±0.52	88.57±0.62	93.03±0.29	81.37±0.73	93.26±0.33
BEiT-3	CE+SupCon	84.93±0.74	93.36±0.34	67.58±0.86	86.10±0.57	91.04±0.55	95.97±0.24	85.72±0.64	95.33±0.29
BEiT-3	CLCE (this work)	**87.00±0.70**	**93.77±0.36**	**69.87±0.91**	**87.06±0.52**	**92.35±0.53**	**96.78±0.23**	**87.24±0.62**	**96.09±0.29**

Table 5.2: Comparison to baselines on the few-shot learning setting. Average few-shot classification accuracies (%) with 95% confidence intervals on test splits of four few-shot learning datasets.

label as the anchor are considered positive, while the rest are viewed as negative. The encoder output, represented by $x_i = Enc(r_i)$, is used to calculate the contrastive loss. In contrast, the output from the classification head, denoted as $z_i = Head(Enc(r_i))$, is used for the CE. We have incorporated L2 normalization on encoder outputs, a strategy demonstrated to enhance performance significantly [235].

5.3.2 Evaluation

We evaluate our proposed approach, CLCE, on image classification in two settings: few-shot learning and transfer learning. We also conduct several analytical experiments. For CLCE experiments, a grid-based hyperparameter search is conducted on the validation set. Optimal settings ($\tau = 0.5$ and $\lambda = 0.9$) are employed because they consistently yield the highest validation accuracies. For all experiments, we use the official train/test splits and report the mean Top-1 test accuracy across at least three distinct initializations.

We employ representative models from two categories of architectures – BEiT-3/MAE/ViT base [115, 224, 252] (transformers based models), and ResNet-101 [81] (convolutional neural network). While new state-of-the-art models are continuously emerging (e.g. DINOv2 [175]), our focus is not on the specific choice of architecture. Instead, we aim to show that CLCE is model-agnostic by demonstrating performance gains with two very different and widely used architectures, as well as show it can be trained and deployed in hardware-constrained settings.

Few-shot Learning

We evaluate our proposed CLCE in the few-shot learning setting, i.e. each test run comprises 3,000 randomly sampled tasks, and we report median Top-1 accuracy with a 95% confidence interval across three runs, maintaining a consistent query shot count of 15. Four prominent benchmarks are used for evaluation: CIFAR-FS [18], FC100 [177], miniImageNet [244], and tieredImageNet [199]. We follow established splitting protocols for a fair comparison [18, 177,

Model	Loss	CIFAR-100	CUB-200	Caltech-256	Oxford-Flowers	Oxford-Pets	iNat2017	Places365	ImageNet-1k
ResNet-101	CE	96.27	84.62	81.38	95.71	93.24	66.11	54.73	78.70
ResNet-101	H-SCL [103]	92.78	77.14	78.64	92.34	92.58	63.14	52.02	77.10
ResNet-101	CE+SupCon	96.31	84.70	81.61	95.73	93.49	66.90	55.41	79.03
ResNet-101	CLCE (this work)	**96.92**	87.48	85.05	**96.33**	94.21	67.93	57.30	80.16
ViT-B	CE	87.13	76.93	90.92	90.86	93.81	65.26	54.06	77.91
ViT-B	CLCE (this work)	88.53	78.21	92.10	92.04	94.01	71.25	58.70	83.94
MAE	CE	87.67	78.46	91.82	91.67	94.05	70.50	57.90	83.60
MAE	CLCE (this work)	90.29	81.30	93.11	92.82	94.88	71.62	58.40	84.02
BEiT-3	CE	92.96	98.00	98.53	94.94	94.49	72.31	59.81	85.40
BEiT-3	H-SCL [103]	89.50	95.70	96.24	92.60	93.28	68.51	56.66	82.25
BEiT-3	CE+SupCon	92.74	98.06	98.65	94.92	94.77	73.58	60.52	85.70
BEiT-3	CLCE (this work)	93.56	**98.93**	**99.41**	95.43	**95.62**	**75.72**	**62.22**	**86.14**

Table 5.3: Comparison to baselines on transfer learning setting. The results are Top-1 classification accuracies across eight diverse datasets.

196].

Tab. 5.2 shows the performance of BEiT-3 and ResNet-101 models under various methods, including CE, H-SCL [103], and the same weighted combination of CE and state-of-the-art supervised contrastive learning loss (SupCon) [109] as CLCE. The results reveal that our CLCE approach consistently improves classification accuracy over other methods, demonstrating superior generalization with limited training data for each class. Our CLCE enhances models' performance on few-shot datasets, significantly outperforming both CE and CE+SuperCon (paired t-test, $p < 0.01$). In the 1-shot learning context when compared to BEiT-3 trained with CE (BEiT-3-CE), the most remarkable improvement is seen on the FC100 dataset, with accuracy rising by 3.52% through the use of CLCE (BEiT-3-CLCE). Indeed, across all datasets, BEiT-3-CLCE shows an average accuracy improvement of 2.7%. For 5-shot learning, the average improvements across the datasets are 1.4% in accuracy for BEiT-3-CLCE, demonstrating CLCE's effectiveness in scenarios with fewer positive samples per class and its ability to yield consistent and reliable results, evident in the tighter confidence intervals for Top-1 accuracy. As for ResNet-101, CLCE (ResNet-101-CLCE) demonstrates even more significant improvements over both CE and CE+SupCon. The enhancement is especially remarkable in the case of tieredImagenet, where ResNet-101-CLCE achieves increases of 16.68% over ResNet-101-CE and 14.17% over ResNet-101-CE+SupCon in 1-shot learning. For 5-shot learning, the improvements are 16.9% and 13.36%, respectively. On average, ResNet-101-CLCE achieves a 9.82% improvement in 1-shot and an 8.71% improvement in 5-shot settings over the ResNet-101-CE. Lastly, H-SCL [103] underperforms compared to CE at a batch size of 128. This highlights contrastive learning's limitation of needing very large batch sizes for better performance than CE, evident in ResNet-101 and BEiT-3 models. Overall, these outcomes underline the efficacy of our proposed CLCE approach and CLCE's broad applicability across different model architectures for few-shot learning tasks.

Transfer Learning

We now assess the transfer learning performance of our proposed CLCE. Here, adhering to the widely accepted paradigm for achieving state-of-the-art results, models are initialized with publicly-available weights from pretraining on ImageNet-21k [45] since they are state-of-the-art, and are fine-tuned on smaller datasets using our new loss function. We leverage 8 datasets: CIFAR-100 [112], CUB-200-2011 [245], Caltech-256 [72], Oxford 102 Flowers [168], Oxford-IIIT Pets [179], iNaturalist 2017 [240], Places365 [291], and ImageNet-1k [45]. We adhere to official train/test splits and report mean Top-1 test accuracy over three different initializations.

Tab. 5.3 presents the results of transfer learning, which offers further evidence of the effectiveness of our proposed CLCE approach beyond few-shot scenarios. When applied to four state-of-the-art image models, including BEiT-3, ResNet-101, ViT-B and MAE, our proposed CLCE approach consistently surpasses other methods, including the standard CE, H-SCL [103] and the same weighted combination of CE and SupCon loss as CLCE. A paired t-test confirms these improvements as statistically significant ($p < 0.05$). While the increase in performance with BEiT-3-CLCE over the BEiT-3-CE baseline is modest in some cases, such as the rise from 98.00% (BEiT-3-CE) to 98.93% (BEiT-3-CLCE) on CUB-200, it shows significant enhancements in challenging datasets with a higher level of class diversity. A notable example is iNaturalist2017, which has 5089 different classes, where CLCE leads to a marked improvement in accuracy from 72.31% to 75.72%. This substantial increase suggests that CLCE's benefits are more pronounced in more varied datasets. In the case of ImageNet-1k, accuracy increased from 85.40% (BEiT-3-CE) to 86.14% (BEiT-3-CLCE), setting a new state-of-the-art for base models (88 million parameters) [2]. We observe similar improvements in other transformer-based models, such as ViT and MAE. The use of CLCE in fine-tuning ResNet-101 also resulted in significant performance gains, particularly in the Caltech-256 dataset. Here, the model's accuracy increases from 81.38% (ResNet-101-CE) to 85.05% (ResNet-101-CLCE). Compared to ResNet-101-CE, there has been an average increase in accuracy of 1.83% for ResNet-101-CLCE. Furthermore, H-SCL [103] yields inferior results compared to CE, mirroring the result observed in few-shot scenarios. Overall, the consistent achievement of high accuracies across diverse datasets using models fine-tuned with CLCE, especially ResNet-101 and BEiT-3, underscores the effectiveness of CLCE in improving model performance. Remarkably, this is achieved without resorting to specialized architectures, extra data, or heightened computational requirements, thereby establishing CLCE as a powerful alternative to traditional CE.

Reducing Batch Size Dependency

We evaluate the effect of batch size on the performance, specifically comparing our CLCE approach with CE and SupCon [109]. The results, as detailed in Tab. 5.4, indicate that SupCon's

[2] `https://paperswithcode.com/sota/image-classification-on-imagenet`

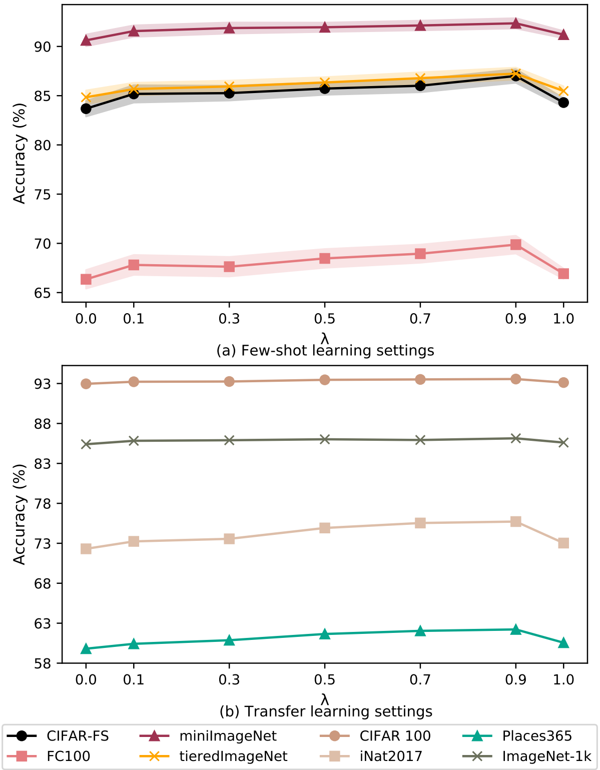


Figure 5.5: Evaluation of the impact of the λ hyperparameter. Results on eight tested datasets with λ values ranging from $\{0, 0.1, 0.3, 0.5, 0.7, 0.9, 1.0\}$. The numerical details for these figures are provided in the supplementary material.

performance is sensitive to batch size variations, a limitation not observed with CE. Particularly, SupCon shows inferior performance compared to CE with the commonly used batch size of 64 on both tested datasets. Even when the batch size is increased to 128, SupCon continues to underperform relative to CE. In our experiments, SupCon generally needs a batch size exceeding 512 to outperform CE, a requirement that is impractical for most single-GPU setups. This scenario mirrors the results of H-SCL [103] in the context of few-shot and transfer learning. In contrast, CLCE not only surpasses CE performance on the iNat2017 dataset with a 1.41% accuracy improvement with batch size of 64 but also demonstrates an even more performance gain of 3.52% in accuracy with batch size of 128. Thus, our CLCE approach significantly mitigates the dependency on large batch sizes typically associated with contrastive learning approaches like SupCon and H-SCL. The reduction in dependency on large batch sizes greatly enhances the adaptability and effectiveness of CLCE in diverse computational settings, such as environments with budget GPUs equipped with 12 GB of memory.

Moreover, gradient accumulation is commonly used in cross-entropy loss to achieve a similar effect when requiring large batch sizes. However, gradient accumulation is very challenging in contrastive learning due to the need to ensure that the accumulated gradients accurately reflect the contrastive nature of the task, particularly in maintaining the integrity of positive and negative pair distributions. This also increases the complexity of maintaining effective sampling

Loss	Batch Size	CIFAR-FS	iNat2017
CE	64	83.68	72.31
CE	128	83.39	72.20
SupCon [109]	64	80.31	69.05
SupCon [109]	128	82.17	69.93
CLCE (this work)	64	84.59	73.72
CLCE (this work)	128	87.00	75.72

Table 5.4: Impact of different batch size. Performance of BEiT-3 base model when trained on CIFAR-FS and iNat2017 datasets. “CE" denotes cross-entropy loss. “SupCon" denotes supervised contrastive learning loss. “CLCE" denotes our proposed joint loss.

CE	CL	HNM	CIFAR-FS	iNat2017
✓			83.68	72.31
✓	✓		84.85	73.53
✓	✓	✓	87.00	75.72

Table 5.5: Results on CIFAR-FS and iNat2017 when training BEiT-3 base model using ablated versions of our CLCE. “CE" denotes cross-entropy loss. “CL" refers to our proposed label-aware contrastive learning, and “HNM" refers to hard negative mining.

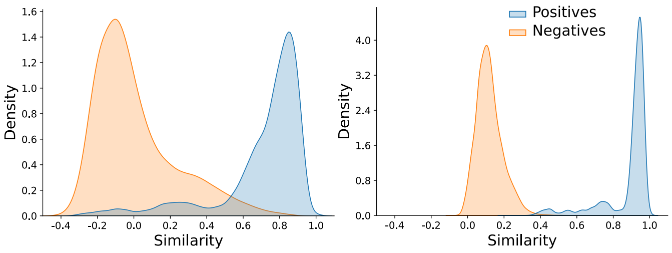
strategies which could vary among datasets, in pairs or triplets across accumulation steps. Thus, gradient accumulation is an inadequate method for overcoming the dependency on large batch sizes. CLCE, on the other hand, offers a more efficient and effective solution.

Optimizing λ: Bridging CE and LACLN

Our proposed CLCE incorporates a hyperparameter, λ, to control the contributions of the CE term and the proposed LACLN term, as shown in Eq. 5.2. To understand the influence of λ, we evaluate its effect on classification accuracy in few-shot learning and transfer learning. Fig. 5.5 presents the test accuracy for varying values of λ. Our experiments reveal a consistent trend: as the weight assigned to the LACLN term (λ) increases, performance progressively improves across all tested datasets, peaking at $\lambda = 0.9$. For instance, this optimal setting yields an average performance boost of 2.14% and 2.74% over the exclusive use of either the LACLN or CE term on four few-shot datasets. This trend also manifests in transfer learning settings, highlighting the complementary nature of CE and LACLN. Thus, optimizing this balance is crucial for maximizing performance with CLCE.

Ablation Study

We conducted an ablation study on the CIFAR-FS and iNat2017 datasets to evaluate the contributions of two key components in our proposed loss: the proposed label-aware contrastive



(a) CE for 'Tulips'

(b) CLCE for 'Tulips'

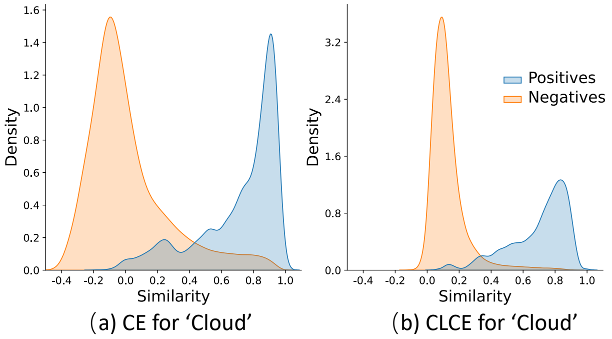
Figure 5.6: Plot of cosine similarity distribution across the "tulips" class from CIFAR-100. Blue represents similarities of positive samples, while orange represents similarities of negative samples.

learning loss without hard negative mining (CL), and the proposed hard negative mining strategy (HNM), as presented in Tab. 5.5. Across both tested datasets, integrating CL with CE is essential for achieving better performance than the CE—e.g. on the CIFAR-FS dataset, there is a notable performance increase of 1.17%. Meanwhile, the integration of our proposed HNM is critical for CLCE's enhanced performance, representing one of the main contributions of this paper. For example, it yields a gain of 2.19% accuracy on the iNat2017 dataset compared to the variant of CLCE without HNM. Hence, we conclude that both components are important and complementary.

Embedding Quality Analysis

We perform a thorough evaluation focusing on the geometric characteristics of the generated representation spaces. We hypothesize that our CLCE enhances the quality of embeddings, thereby sharpening class distinction and improving performance. To elaborate, we examine the CE embeddings and CLCE embeddings produced by the BEiT-3 base model. Specifically, we evaluate two key aspects: (1) Distributions of cosine similarities between image pairs. This assessment provides insights into how well the model differentiates between classes in the embedding space. (2) Visualization of the embedding space using the t-SNE algorithm [239]. This visualization allows us to observe the separation or clustering of data points belonging to different classes. (3) We employ the Isotropy Score as defined by [163] to evaluate the quality of produced embeddings. The Isotropy Score measures the distribution of data in the embedding space and serves as a metric for the quality of the produced embeddings. Historically, isotropy has served as an evaluation metric for representation quality [9]. This is based on the premise that widely distributed representations across different classes in the embedding space facilitate better distinction between them.

We present the pairwise cosine similarity distributions of CE and CLCE embeddings in Figs. 5.6 and 5.7. Specifically, we randomly select the "tulips" and "cloud" classes from CIFAR-100 to compute cosine similarities for positive (same class) and negative pairs (different classes). Observations from these plots reveal that the CLCE embeddings demonstrate superior separation



(a) CE for 'Cloud'

(b) CLCE for 'Cloud'

Figure 5.7: Plot of cosine similarity distribution across the "cloud" class from CIFAR-100. Blue represents similarities of positive samples, while orange represents similarities of negative samples.

Model	iNaturalist2017	Imagenet-1k	Places365
BEiT3-CE	0.32	0.27	0.34
BEiT3-CLCE	0.98	0.92	0.93

Table 5.6: Comparison of Isotropy Score across three datasets for BEiT-3-CE and BEiT-3-CLCE. A higher value is better. A higher Isotropy Score indicates better isotropy and generalizability.

between classes and less overlap between positive and negative samples compared to CE.

In Fig. 5.8, the t-SNE visualization of the embedding space for CE and CLCE across twenty CIFAR-100 classes. The CE embeddings (Fig. 5.8a) display instances where the same class nodes are relatively closely packed but also reveal many outliers. This suggests a reduced discriminative capability. On the contrary, CLCE embeddings (Fig. 5.8b) display more separated and compact class clusters, suggesting improved discriminative capabilities.

Formally, we calculate the quantitative Isotropy Score (IS) [163], which is defined as follows:

$$IS(\mathcal{V}) = \frac{max_{c \subset C} \sum_{v \subset V} \exp{(C^T V)}}{min_{c \subset C} \sum_{v \subset V} \exp{(C^T V)}} \tag{5.5}$$

where V is a set of vectors, C is the set of all possible unit vectors (i.e., any c so that $||c|| = 1$) in the embedding space. In practice, C is approximated by the eigenvector set of $V^T V$ (V are the stacked embeddings of v). The larger the IS value, the more isotropic an embedding space is (i.e., a perfectly isotropic space obtains an IS score of 1).

These observations indicate that the proposed CLCE approach restructures the embedding space to enhance class distinction, addressing the generalization limitation of the CE. This enhancement is particularly effective for few-shot scenarios, where limited labelled data requires the model to rely more on high-quality, discriminative representations.

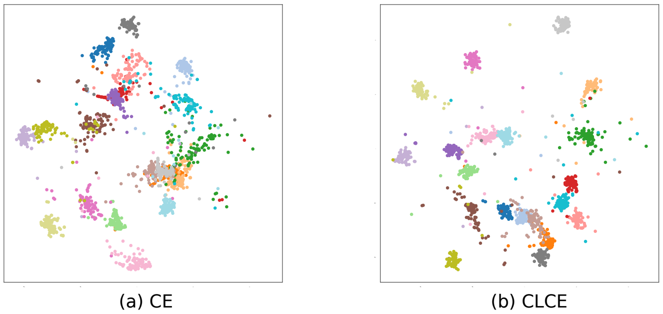

(a) CE
(b) CLCE

Figure 5.8: Embedding Space Visualization for CE vs. CLCE, over twenty CIFAR-100 test set classes using t-SNE. Each dot represents a sample, with distinct colors indicating different label classes.

5.3.3 Discussion and Summary

Limitations. While our CLCE approach advances the state-of-the-art, it still has certain limitations. Firstly, CLCE shows increased performance with larger batch sizes. As Table 5.4 illustrates, CLCE surpasses CE in accuracy in few-shot and transfer learning scenarios at a batch size of 64, with further improvements observed at larger batch sizes. Secondly, our approach applies hard negative mining solely to the contrastive learning component and not to the CE component. This is due to differing implementations of hard negative mining in each loss. In cross-entropy, hard negatives are identified based on loss values, necessitating a unique strategy that might interfere with the existing sampling process in contrastive learning and potentially cause conflicting outcomes. Additionally, the divergent goals of cross-entropy and contrastive learning, where the former focuses on minimizing the discrepancy between predicted and true distributions and the latter emphasizes embedding similarities, complicate the use of a unified hard negative mining approach.

Summary. In this Section, we proposed a approach for training image models, denoted CLCE. CLCE combines label-aware contrastive learning with hard negative mining and CE, to address the shortcomings of CE and existing contrastive learning methods. Our empirical results demonstrate that CLCE consistently outperforms traditional CE and prior contrastive learning approaches, both in few-shot learning and transfer learning settings. Furthermore, CLCE offers an effective solution for researchers and developers who can only access commodity GPU hardware, as CLCE maintains its effectiveness when working with smaller batch sizes that can be loaded onto cheaper GPU cards with less on-board memory. To summarize, our comprehensive investigations and robust empirical evidence compellingly substantiate our methodological decisions, demonstrating that CLCE is a superior alternative to CE for enhancing the performance of image models in image classification. This supports our thesis statement by showcasing how improved alignment and efficiency in multimodal learning frameworks can lead to significant performance gains across various tasks.

5.4 Elucidating and Overcoming the Challenges of Label Noise in Supervised Contrastive Learning

After enhancing contrastive learning by incorporating hard negative mining and cross-entropy loss into the overall objective in Section 5.3, we further explore the factors influencing its performance.

5.4.1 Recap

Contrastive methods achieve excellent performance on self-supervised learning [64, 82, 186]. They produce latent representations that excel at many downstream tasks, from image recognition and object detection to visual tracking and text matching [74, 232]. *Supervised* contrastive learning (SCL) utilizes label information to improve representation learning, encouraging closer distances between same-class samples (positive pairs) and greater distances for different-class samples (negative pairs). SCL outperforms traditional methods for pre-training that employ a cross-entropy loss [74, 109, 128].

5.4.2 Introduction

Indeed, the effectiveness of SCL depends on the correctness of the labels used for identifying image pairs to contrast. Human-labelling errors introduce erroneous positive and negative pairings, compromising the integrity of learned representations [128]. Even widely-used datasets exhibit significant numbers of mislabeled images—for example, the ImageNet-1K validation set has 5.83% of images wrongly labeled [169]. The impact of noisy labels on supervised learning has been extensively researched [53, 77, 219]. However, the extent and manner in which human-labelling errors influence SCL remains under-explored. As SCL emerges as a compelling alternative to the cross-entropy loss, it becomes increasingly important to investigate human-labelling errors specifically in the context of SCL.

In this section, we first analyze how human-labelling errors affect SCL and how they differ from supervised learning with cross-entropy loss. As shown on the left side of Fig. 5.9 and detailed in Sec. 5.4.4, labelling errors, regardless of being positive or negative, negatively impact supervised learning with cross-entropy loss during training. Existing noise-mitigation methods, based on SCL or cross-entropy loss approaches [94, 128], aim to eliminate labelling errors from training samples. However, we argue that such strategies, often detrimental to SCL, compromise the quality of learned representations. The middle of Fig. 5.9 shows that labelling errors do not always adversely affect SCL, with their removal potentially reducing training sample size and lowering overall performance. Unlike in cross-entropy methods, labelling errors exhibit more complex dynamics in how they affect learning signals in SCL. Both correctly and incorrectly labeled images can generate correct or incorrect learning signals, depending on the images they

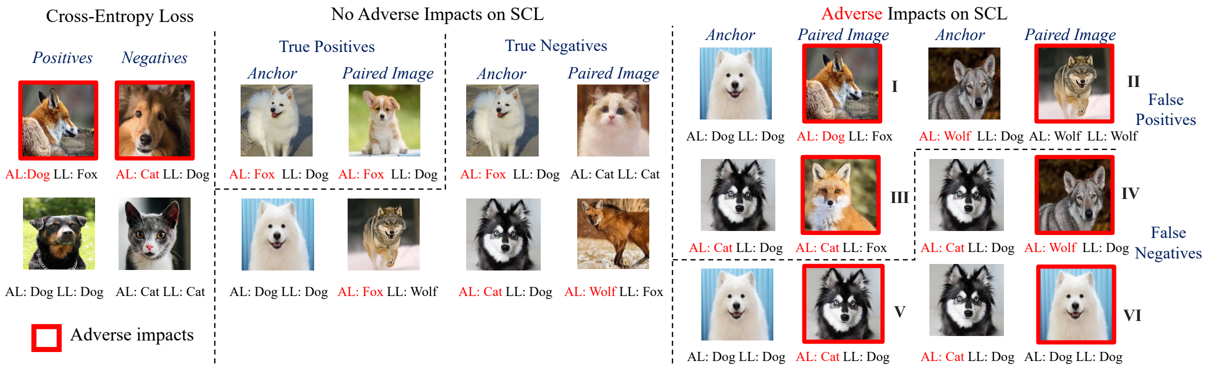


Figure 5.9: Comparison between impacts of labelling errors on different learning approaches. AL represents 'Assigned Label' and LL represents 'Latent Label'. Those marked red in AL represent human-labelling errors. It is important to note that as long as a pair shares the same latent label, there are no adverse impacts on positive pairs. Similarly, if the latent labels differ, negative pairs remain unaffected.

are paired with. Our analysis in Sec. 5.4.4 reveals that nearly 99% of incorrect learning signals in SCL are due to mislabeled positive samples. This highlights the need for a tailored SCL strategy that effectively addresses human-labelling errors without sacrificing performance.

Although human-labelling errors are prevalent in image datasets, existing noise-mitigation methods predominantly focus on synthetic labelling errors, showing good performance when noise is intentionally added at levels ranging from 40% to 80% [94, 128, 171, 206]. Yet, in scenarios with realistic human-labelling errors, such as those found in ImageNet-1K (with a noise rate of approximately 5.83% [169]), these methods underperform, especially against cross-entropy based methods, primarily due to overfitting [169, 219, 268, 281]. For example, methods that rely on assigning confidence values to pairs and prioritizing learning from those deemed confident [94, 128] risk overfitting to incorrect labels and neglect a significant portion of training data, an issue exacerbated in datasets with many similar classes [219, 268]. Additionally, existing noise-mitigating methods also introduce significant computational complexity and overhead. For example, Sel-CL [128] is challenging to apply on large datasets since it uses the *k*-NN algorithm to create pseudo-labels.

Importantly, in Sec. 5.4.4, we reveal significant differences between the distributions of synthetic and human-labelling errors, emphasizing the need for tailored mitigation strategies. Our empirical analysis indicates that human-labelling errors often stem from high visual similarity between assigned and actual classes, leading to notable overlaps in the representation distributions of correctly and incorrectly labeled samples. This indicates that human-mislabeled samples are often indistinguishable in the representation space to samples sharing the same assigned label; this contrasts with synthetic label errors that are more arbitrary. Thus, there is an urgent need for methods that not only mitigate the impact of human-labelling errors in SCL on widely-used human-annotated datasets like ImageNet-1K but also preserve computational efficiency.

Based on our analysis and existing research gaps, we propose a novel SCL objective with robustness to human-labelling errors that emphasises true positives, which come from the same latent class but are are far apart in feature space (Sec. 5.4.5). In contrast, existing noise-mitigating methods tuned for synthetic noise[7, 85] typically assign greater weight to confident pairs that are closely positioned, despite the high likelihood of these pairs being false positives. Furthermore, based on the established benefits of utilizing ‘hard’ negative samples [171, 203, 206], we hypothesise that *true positives originating from the same latent class, yet positioned distantly, are important for enhancing the quality of learned representations*.

In summary, our main contributions are:

- We present an in-depth analysis elucidating the impact of human-mislabeled samples on supervised contrastive learning (SCL) and offer strategies to effectively mitigate these issues (Sec. 5.4.4). This analysis aligns with our thesis statement by addressing critical factors that influence model performance and reliability.

- We introduce a novel SCL objective, *SCL-RHE*, the first to specifically address human-labeling errors by focusing on false positives within SCL (Sec. 5.4.5). Unlike previous works [169, 219, 268, 281], SCL-RHE not only demonstrates state-of-the-art performance on widely used image datasets and avoids overfitting, but also maintains efficiency without adding extra computational overhead. This innovation supports our thesis statement by enhancing the accuracy and efficiency of multimodal learning frameworks.

- We demonstrate the broad applicability of the proposed SCL-RHE objective across two distinct learning scenarios: training from scratch and transfer learning. SCL-RHE outperforms previous state-of-the-art SCL and noise-mitigation methods, achieving higher Top-1 accuracy on ten tested datasets in both scenarios, as detailed in Sec. 5.4.6. Notably, SCL-RHE sets a new state-of-the-art for base models (with 88M parameters) on ImageNet-1K. These results validate our thesis claim of improved alignment and performance in multimodal learning frameworks.

- The original material in this section has been accepted for presentation at the 18th European Conference on Computer Vision (ECCV), a Core ranking A* conference with an h5-index of 238.

Based on this analysis, we propose a novel sampling strategy that emphasises true positives, which come from the same latent class but are distantly placed, and true negatives with similar representations (Section. 5.4.5). In contrast, existing noise-robust methods [7, 85] typically assign greater weight to confident pairs that are closely positioned, despite the high likelihood of these pairs being false positives. Furthermore, based on the established benefits of utilizing ‘hard’ negative samples [171, 203, 206], we hypothesise that *true positives originating from the same latent class, yet positioned distantly, could be important for enhancing the quality of learned representations*. As a result, we strategically reduce the risk of an incorrect decision boundary, as shown in Fig. 5.9. In summary, our main contributions in this Section are:

- **Insights into the impact of noisy labels on supervised contrastive learning**: We present an in-depth analysis in Section. 5.4.4, elucidating the impact of mislabeled samples on SCL and offering strategies for their effective mitigation.

- **A novel technique for efficient, noise-robust contrastive learning**: We propose *D-SCL* in Section. 5.4.5, a novel SCL objective that is the first to remove bias due to misannotated labels within a supervised contrastive framework. Unlike previous works [27, 94, 128, 289], D-SCL does not introduce extra computational overhead and is suitable for general image benchmarks with large image models, thereby optimizing performance without sacrificing processing speed or resource utilization.

- **State-of-the-art performance**: Compared to traditional pre-training using cross-entropy, D-SCL achieves significant gains in Top-1 accuracy across multiple datasets including iNat2017 [240], ImageNet [45], and others (Section. 5.4.6). Furthermore, when using existing pre-trained weights, D-SCL demonstrates superior performance on transfer learning, outperforming existing methods such as Sel-CL [128] and setting a new state-of-the-art for base models (with 88M parameters) on ImageNet-1K. It shows a noticeable increase of 3.4% in performance on the iNat2017 dataset.

5.4.3 Setup for Supervised Contrastive Learning

We begin by discussing the fundamentals of contrastive representation learning. Here, the objective is to contrast pairs of data points that are semantically similar (positive pairs) against those that are dissimilar (negative pairs). Mathematically, given a data distribution $p(x)$ over $\mathscr{X}$, the goal is to learn an embedding $f : \mathscr{X} \to \mathbb{R}^d$ such that similar pairs (x, x^+) are close in the feature space, while dissimilar pairs (x, x^-) are more distant. In unsupervised learning, for each training datum x, the selection of x^+ and x^- is dependent on x. Typically, one positive example x^+ is generated through data augmentations and N negative examples x^-. The contrastive loss, named InfoNCE or the N-pair loss [76, 172, 217], is then defined as

$$\mathscr{L}_{\text{NCE}} = \mathbb{E}_{\substack{x \\ x^+ \\ \{x_i^-\}_{i=1}^N}} \left[-\log \frac{e^{f(x)^T f(x^+)}}{e^{f(x)^T f(x^+)} + \sum_{i=1}^N e^{f(x)^T f(x_i^-)}} \right] \tag{5.6}$$

Here, the expectation computes the average loss across all possible choices of positive and negative samples within the dataset. In practice, during a training iteration, one typically samples a mini-batch; then, for each data point in it (referred to as an 'anchor'), a positive example is selected—usually an augmented version of the anchor or another instance of the same class—while the rest of the batch is treated as negative examples. This is under the assumption that within the batch, instances of different classes (i.e., all other samples except the positive pair) serve as negatives.

Khosla *et al.* [109] extended this concept to supervised contrastive learning, experimenting with two losses:

$$\mathscr{L}_{\text{in}}^{\text{sup}} = \mathbb{E}_{\substack{x \\ \{x_k^+\}_{k=1}^K \\ \{x_i^-\}_{i=1}^N}} - \log \left\{ \frac{1}{|K|} \sum_{k=1}^K \frac{\exp\big(f(x)^T f\big(x_k^+\big)\big)}{\sum_{k=1}^K e^{f(x)^T f(x_k^+)} + \sum_{i=1}^N e^{f(x)^T f(x_i^-)}} \right\} \tag{5.7}$$

$$\mathscr{L}_{\text{out}}^{\text{sup}} = \mathbb{E}_{\substack{x \\ \{x_k^+\}_{k=1}^K \\ \{x_i^-\}_{i=1}^N}} \frac{-1}{|K|} \sum_{k=1}^K \left[\log \left\{ \frac{\exp\big(f(x)^T f\big(x_k^+\big)\big)}{\sum_{k=1}^K e^{f(x)^T f(x_k^+)} + \sum_{i=1}^N e^{f(x)^T f(x_i^-)}} \right\} \right] \tag{5.8}$$

Here k indexes a set of K positive samples, i.e. images x_k^+ of the same class as x. It is unnecessary to be concerned about negative values in the contrastive learning objective. This is due to the fact that each term in both the numerator and denominator incorporates the exponential of a similarity score. The exponential function, exp(z), remains positive irrespective of the value of z, thereby ensuring that every term in both the numerator and the denominator is positive.

We focus on improving these objectives, making them more robust to labeling errors in Section. 5.4.5.

5.4.4 Uniqueness of Human-labelling Errors and Their Impact on SCL

In this section, we show that human-labelling errors and synthetic label errors exhibit distinct characteristics. Human-labelling errors arise from the high visual similarity between the sample and its assigned class, making it challenging for humans to differentiate them accurately. In contrast, synthetic label errors are generated randomly and lack this similarity [94, 128]. This distinction underlines the need for a method specifically tailored to address the unique challenges human-mislabelled samples pose to supervised contrastive learning (SCL). We further illustrate the specific impact of these errors on SCL, distinct from their effect on supervised learning with cross-entropy, by analyzing various scenarios of mislabelling within SCL and assessing their adverse effects.

Definitions. We define the *latent label* of an image as being its true category (e.g. the latent label of an image of a cat would be 'cat'). The term *assigned label* refers to the class that a human annotator has assigned to an image (hopefully—but not always—matching the latent label). Given a pair of images, we define a *false positive* as being when an annotator has erroneously grouped those images under the same assigned label, even though their latent labels are different. A *false negative* is when two images sharing the same latent label mistakenly have different assigned labels. We define *true positive* and true negative pairs analogously. Lastly, we define *easy positives* as pairs of images that share the same assigned labels and have highly similar embeddings.

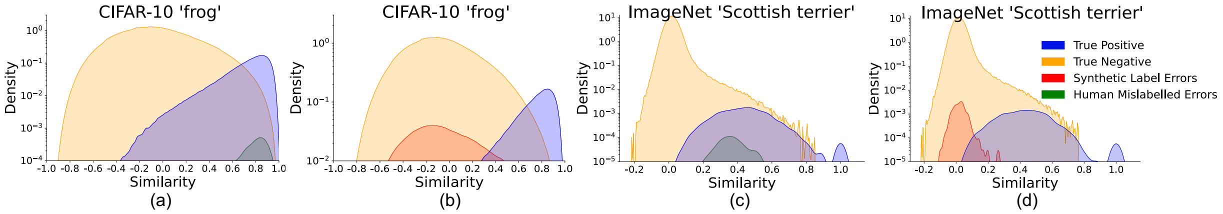


Figure 5.10: Figures (a) and (c) display the log-scaled distribution of cosine similarities for various pair types, including true positive pairs, true negative pairs, and human-labelling errors, on the CIFAR-10 and ImageNet-1k datasets, respectively. Conversely, figures (b) and (d) present analogous data, focusing instead on synthetic label errors.

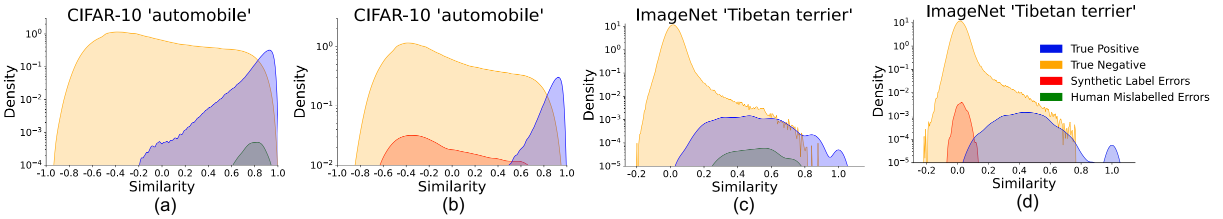


Figure 5.11: Figures (a) and (c) display the log-scaled distribution of cosine similarities for various pair types, including true positive pairs, true negative pairs, and human-labelling errors, on the CIFAR-10 and ImageNet-1k datasets, respectively. Conversely, figures (b) and (d) present analogous data, focusing instead on synthetic label errors.

The Differences Between Human-Labelling Errors and Synthetic Label Errors

We begin our analysis with the question: *What distinguishes human-labelling errors from synthetic label errors?* We pretrained ViT-base models on the CIFAR-10 and ImageNet-1k datasets individually, utilizing the SCL objective as defined in Eq. 5.6. Then, we conduct a similarity analysis of the resulting features for different types of label errors within the context of contrastive learning. Specifically, we use the consensus among annotators from [169] to identify human-mislabelled samples, and synthetic errors are produced by randomly altering 20% of the labels to different classes. In Fig. 5.10, we plot the similarity distributions for various pair types across two image classes, contrasting human-labelling errors with synthetic label errors. As we can see from Fig. 5.10, there is a significant overlap in the similarity distributions of true positives and human-labelling errors (false positives), indicating a high similarity in their embeddings, which is notably larger than that observed with true negatives. In contrast, the overlap between true positives and synthetic label errors (false positives) is not obvious. Fig. 5.11 (a) and (c) illustrate the similarity maps for true positive pairs, true negative pairs, and human-labeling errors with true positive pairs within the CIFAR-10 and ImageNet datasets, respectively, while Fig. 5.11 (b) and (d) present the similarity maps for true positive pairs, true negative pairs, and synthetic errors with true positive pairs for the same datasets. Across both datasets, it is evident that human-labeling errors exhibit a significantly higher similarity with true positive pairs compared to negative ones. This result demonstrates that human-labelling errors primarily arise due to high visual similarity between the assigned and latent class, unlike synthetic label errors.

This empirical finding of a small overlap between true positives and synthetic label errors explains the effectiveness of synthetic noise-mitigating methods such as Sel-CL [128] and TCL [94], which excel by allocating greater weight to confident pairs that are closely aligned. However, it also underscores the limitation of applying the same strategy to address the impact of human-labelling errors, which are close to true positives, serving as a primary motivation for this work.

Furthermore, the significant overlap between true positives and human-labelling errors reveals that human-labelling errors are 'easy positives', indicating that the embeddings of positive samples are closely clustered in the representation space. This insight informs our strategy of reducing the weighting of easy positives as an effective means to mitigate the impact of false positives resulting from human-labelling errors.

Impacts of Human-Labelling Errors on SCL

Based on the conclusions of Sec. 5.4.4, we now know we need to focus on 'easy positives'. It also raises the question: *What is the impact of human-mislabelled samples when they appear as negatives?* Therefore, in this section, we give a deeper analysis of the interaction of human-labelling errors and SCL by looking at the probability of false positives and false negatives during training.

Let 'A' represent an anchor image and 'B' represent another paired image. If 'A' and 'B' are assigned the same label, they are a false positive if: 'A' is correctly labeled while 'B' is not (I in Fig.5.9(a)); *or* 'A' is mislabelled while 'B' is correctly labeled (II in Fig.5.9(a)); *or* 'A' and 'B' are mislabelled and do not belong to the same latent class (III in Fig.5.9(a)). Conversely, if 'A' and 'B' are assigned different labels, they are a false negative if both 'A' and 'B' are mislabelled but actually belong to the same latent class (IV in Fig.5.9(a)); *or* 'A' is correctly labeled and 'B' is not, yet 'B' belongs to the same latent class as 'A' (V in Fig.5.9(a)); *or* 'A' is mislabelled and 'B' is not, with 'A' being of the same latent class as 'B' (VI in Fig.5.9(a)).

Assuming human-labelling errors rate is τ, we can derive the probability of mislabelled data appearing as false negatives, $P_{FN} = \frac{\tau^2}{(C-2)^2} + \frac{2\tau-2\tau^2}{C-1}$, and as false positives, $P_{FP} = 2\tau - \tau^2 - \frac{\tau^2}{(C-1)^2}$. C is the number of classes and τ is the error rate. Since τ is small, terms with τ^2 are negligible. Therefore, $P_{FN} \approx \frac{2\tau}{C-1}$ and $P_{FP} \approx 2\tau$. As C increases, P_{FN} tends to zero, while P_{FP} remains constant. Therefore, with many classes and a small error rate, $P_{FP} \gg P_{FN}$. For example, if there are 200 classes in a dataset with a 5% error rate, we would expect wrong learning signals from false positives and false negatives with probabilities of 9.75% and 0.05%, respectively.

Additionally, we substantiate this finding with empirical evidence by quantifying the incorrect learning signals from human-labelling errors during training on the original CIFAR-100 and ImageNet-1K datasets separately, based on human-labelling errors provided by [169]. When training on the CIFAR-100 dataset (comprising 100 classes), we found that 99.04% of the incorrect learning signals were caused by human-labelling errors, from false positives, with only approximately 0.96% stemming from negatives. Similarly, when training on the ImageNet-1K dataset (comprising 1000 classes), we discovered that 99.91% of the incorrect learning signals were due to human-labelling errors, primarily from false positives, with only about 0.09% arising from negatives. We provide these rates for other datasets in the supplementary material.

Overall, both theoretical and empirical results show that when tackling labelling errors in contrastive learning, we can largely ignore false negatives due to their very low rate of occurrence, and focus on easy positives. These observations motivate our proposed method in Section 5.4.5, which incorporates less weighting on easy positives and reduces the wrong learning signal caused by human-labelling errors.

5.4.5 SCL with Robustness to Human-Labelling Errors

In this section, we describe our approach to mitigate the impacts caused by human-labelling errors in positive pairs and how this fits into an overall contrastive learning objective. In Sec. 5.4.4, we noted that the most significant impact of the mislabeled samples arises when they are incorporated into the positive set and exhibit high similarity to the anchor (i.e. easy positives). Our method therefore adheres to two key principles (Fig. 5.9b): (**P1**) it should ensure that the *latent* class of positive samples matches that of the anchor [40, 106, 186, 204]; and (**P2**) it should deprioritize easy positives, i.e. those currently embedded near the anchor. By reducing the weighting of easy positives, we minimize the effect of incorrect learning signals from false positive pairs. The model is also forced to recognize and encode deeper similarities that are not immediately apparent, improving its discriminative ability.

Human-Labelling Errors in the SCL Objective

Khosla *et al.* [109] argued that $\mathscr{L}_{\text{out}}^{\text{sup}}$ is superior to $\mathscr{L}_{\text{in}}^{\text{sup}}$, attributing this to the normalization factor $\frac{1}{|P(i)|}$ in $\mathscr{L}_{\text{out}}^{\text{sup}}$ that mitigates bias within batches. Although $\mathscr{L}_{\text{in}}^{\text{sup}}$ incorporates the same factor, its placement inside the logarithm reduces its impact to a mere additive constant, not influencing the gradient and leaving the model more prone to this bias.

We instead introduce a modified $\mathscr{L}_{\text{in}}^{\text{sup}}$ that directly reduces bias due to mislabeling, and outperforms $\mathscr{L}_{\text{out}}^{\text{sup}}$.

We begin with a modified formulation of $\mathscr{L}_{\text{in}}^{\text{sup}}$ (Eq. 5.7), that is equivalent up to a constant scale and shift, but will prove easier to adapt:

$$\mathop{\mathbb{E}}_{\substack{x \\ \{x_k^+\}_{k=1}^K \\ \{x_i^-\}_{i=1}^N}} \left[-\log \frac{1}{|K|} \frac{\sum_{k=1}^K e^{f(x)^T f(x_k^+)}}{\sum_{k=1}^K e^{f(x)^T f(x_k^+)} + \sum_{i=1}^N e^{f(x)^T f(x_i^-)}} \right] \tag{5.9}$$

In 5.9, all K samples from the same class within a mini-batch are treated as positive samples for the anchor x.

We now introduce our main technical contribution, which is devising an objective that mitigates human-labelling errors, which consists of modifying Eq. (5.9). As in Sec. 5.4.4, we assume there is set of latent classes $\mathscr{C}$, that encapsulate the semantic content, and hopefully match the assigned labels. Following [8, 40, 203], pairs of images (x, x^+) are supposed to belong to the same latent class c, where $x \in \mathscr{X}$ is drawn from a data distribution $p(x)$. Let τ denote the probability that any sample is mislabeled; we assume this is constant for all x. Since τ is unknown in practice, it must be treated as hyperparameter, or estimated based on previous studies. We also introduce an (unknown) function $z : \mathscr{X} \to \mathscr{C}$ that maps x to its latent class label. Then, $p_x^+ := p\big(x' \,|\, z(x') = z(x)\big)$ is the probability of observing x' as a positive example of x, whereas $p_x^- = p\big(x' \,|\, z(x') \neq z(x)\big)$ is the probability of a negative example. For each image x, the objective (Eq. 5.9) aims to learn a representation $f(x)$ by using positive examples $\{x^+\}_{k=1}^K$ with the same latent class label as x and negative examples $\{x_i^-\}_{i=1}^N$ that belong to different latent classes. Since p is the true data distribution, the ideal loss function to be minimized if the latent labels $z(x)$ were known is:

$$\mathscr{L}_T = \mathbb{E}_{\substack{x \sim p \\ x_k^+ \sim p_x^+ \\ x_i^- \sim p_x^-}} \left[\frac{-1}{|K|} \log \frac{\frac{Q}{K} \sum_{k=1}^K e^{f(x)^T f(x_k^+)}}{\frac{Q}{K} \sum_{k=1}^K e^{f(x)^T f(x_k^+)} + \frac{W}{N} \sum_{i=1}^N e^{f(x)^T f(x_i^-)}} \right] \tag{5.10}$$

We term this loss function the *true label loss*. Here, we have introduced weighting parameters Q and W to help with analysing the impacts of human-labelling errors; when they equal the numbers of positive and negative examples respectively, $\mathscr{L}_T$ reduces to the conventional supervised contrastive loss (5.9). Note that supervised contrastive learning typically assumes p_x^+ and p_x^- can be determined from human annotations (i.e. $z(x)$ yields the *assigned* label of x); however since we consider latent classes instead of assigned classes, we do *not* have access to the true distribution. However, we now show how to approximate this true distribution and improve the overall performance.

Mitigating Human-Labelling Errors

For a given anchor x and its embedding $f(x)$, we now aim to build a distribution q on $\mathscr{X}$ that fulfils the principles **P1** and **P2**. We draw a batch of positive samples $\{x_k^+\}_{k=1}^K$ from q. Ideally we would draw samples from

$$q^+(x^+) := q\big(x^+ \,|\, z(x) = z(x^+)\big) \propto \frac{1}{e^{\beta f(x)^T f(x^+)}} \cdot p_x^+(x^+) \tag{5.11}$$

where $\beta \geq 0$. It is important to note that $q^+(x^+)$ depends on x, although this dependency is not explicitly shown in the notation. The distribution is composed of two factors:

- The event $\{z(x) = z(x^+)\}$ indicates that pairs, (x, x^+), should originate from the same latent class (**P1**); recall $p_x^+(x^+)$ is the true (unknown) positive distribution for anchor x.

- The exponential term increases the probability of sampling hard positives, and decreases that of sampling easy positives (**P2**). This term is an unnormalized von Mises-Fisher density with mean direction $f(x)$ and a concentration parameter β. The concentration parameter β modulates the weighting scheme of q^{+}, specifically augmenting the weights of instances x^{+} that exhibit a lower inner product (i.e. greater dissimilarity) to the anchor x.

The distribution q^{+} fulfils our desired principles of selecting true positives and deprioritizing easy positives. However, we do not have the access to the latent classes, and so cannot directly sample from it. We therefore rewrite it from the perspective of Positive-Unlabeled (PU) learning [40, 52, 55, 203], which will allow us to implement an efficient sampling mechanism. We first define $q^{-}(x^{+}) \propto \frac{1}{e^{\beta f(x)^T f(x^{+})}} \cdot p_x^{-}(x^{+})$. Then, by conditioning on the event $\{z(x) = z(x^{+})\}$, we can write

$$q(x^{+}) := \tau^{+} q^{+}(x^{+}) + \tau^{-} q^{-}(x^{+}) \tag{5.12}$$

$$\Rightarrow \; q^{+}(x^{+}) = \left(q(x^{+}) - \tau^{-} q^{-}(x^{+})\right) / \tau^{+} \tag{5.13}$$

where τ^{+} is the probability that a sample from the data distribution $p(x)$ will have the same latent class as x.

We have now derived an alternative expression (5.13) for the positive sampling distribution q^{+} in terms of q and q^{-}. Sampling directly from q and q^{-} is still not possible; however, we can use importance sampling to approximate the necessary expectations. Specifically, we shall choose positives primarily by sampling an assigned-positive, while occasionally sampling an assigned-negative, with the probability of the latter set to counterbalance the mislabeling rate.

To achieve this we first consider a sufficiently large value for K (i.e. the number of positive samples for the anchor x) in the SCL objective (5.10), while holding the weighting parameter Q fixed. Then, (5.10) becomes:

$$L_T = \mathbb{E}_{\substack{x \sim p \\ x^{-} \sim p_x^{-}}} \left[-\log \frac{1}{|K|} \frac{Q\mathbb{E}_{x^{+} \sim q^{+}}\left[e^{f(x)^T f(x^{+})}\right]}{Q\mathbb{E}_{x^{+} \sim q^{+}}\left[e^{f(x)^T f(x^{+})}\right] + \frac{W}{N}\sum_{i=1}^{N} e^{f(x)^T f(x_i^{-})}} \right] \tag{5.14}$$

By substituting Eq. 5.13 into Eq. 5.14, we obtain an objective that rectifies the impacts from human-labeling errors and also down-weights easy positives:

$$\mathbb{E}_{\substack{x \sim p \\ x^{-} \sim q}} \left[-\log \frac{1}{|K|} \frac{\frac{Q}{\tau^{+}}\left(\mathbb{E}_{x^{+} \sim q}\left[e^{f(x)^T f(x^{+})}\right] - \tau^{-}\mathbb{E}_{v \sim q^{-}}\left[e^{f(x)^T f(v)}\right]\right)}{\frac{Q}{\tau^{+}}\left(\mathbb{E}_{x^{+} \sim q}\left[e^{f(x)^T f(x^{+})}\right] - \tau^{-}\mathbb{E}_{v \sim q^{-}}\left[e^{f(x)^T f(v)}\right]\right) + \frac{W}{N}\sum_{i=1}^{N} e^{f(x)^T f(x_i^{-})}} \right] \tag{5.15}$$

This suggests we only need to approximate the expectations $\mathbb{E}_{x^{+} \sim q}\left[e^{f(x)^T f(x^{+})}\right]$ and $\mathbb{E}_{v \sim q^{-}}\left[e^{f(x)^T f(v)}\right]$ over q and q^{-}, which can be achieved by classical Monte Carlo importance sampling, using samples from p and p^{-}:

$$\mathbb{E}_{x^+\sim q}\left[e^{f(x)^T f\left(x^+\right)}\right]=\mathbb{E}_{x^+\sim p}\left[e^{f(x)^T f(x^+)}q/p\right]=\mathbb{E}_{x^+\sim p}\left[e^{(\beta+1)f(x)^T f(x^+)}/Z(x)\right] \tag{5.16}$$

$$\mathbb{E}_{v\sim q^-}\left[e^{f(x)^T f(v)}\right]=\mathbb{E}_{v\sim p^-}\left[e^{f(x)^T f(v)}q^-/p^-\right]=\mathbb{E}_{v\sim p^-}\left[e^{(\beta+1)f(x)^T f(v)}/Z^-(x)\right] \tag{5.17}$$

where $Z(x)$ and $Z^-(x)$ are the partition functions for q and q^- respectively. Hence, these expectations over p and p^- admit empirical estimates

$$\widehat{Z}(x)=\frac{1}{M}\sum_{i=1}^{M}e^{\beta f(x)^\top f(x_i^+)} \qquad \text{and} \qquad \widehat{Z}^-(x)=\frac{1}{N}\sum_{i=1}^{N}e^{\beta f(x)^\top f(x_i^-)}. \tag{5.18}$$

Mitigating label errors for negatives. Despite the minimal impact of mislabeled samples in negative sets (see Sec. 5.4.4), we extend our mitigation method to these samples to further reduce their adverse effects. Mirroring our strategy for positive samples, the mitigation process for negatives involves constructing a distribution that not only aligns with true negatives but also places greater emphasis on hard negatives. We then use Monte Carlo importance sampling techniques to better estimate the true distribution of latent classes. Full details are given in the supplementary material.

Overall learning objective. Using Eq. 5.15 and incorporating mitigation for negatives, we get our final SCL-RHE loss

$$\mathbb{E}_{\substack{x\sim p\\ x^+\sim q\\ x^-\sim q}}\left[-\log\frac{1}{|K|}\frac{\frac{Q}{\tau^+}\left(\mathbb{E}_{x^+\sim q}\left[e^{f(x)^T f(x^+)}\right]-\tau^-\mathbb{E}_{v\sim q^-}\left[e^{f(x)^T f(v)}\right]\right)}{\frac{Q}{\tau^+}\left(\mathbb{E}_{x^+\sim q}\left[e^{f(x)^T f(x^+)}\right]-\tau^-\mathbb{E}_{v\sim q^-}\left[e^{f(x)^T f(v)}\right]\right)+\frac{W}{\tau^-}\left(\mathbb{E}_{x^-\sim q}\left[e^{f(x)^T f(x^+)}\right]-\tau^+\mathbb{E}_{b\sim q^+}\left[e^{f(x)^T f(b)}\right]\right)}\right] \tag{5.19}$$

This objective has the following desirable properties:

- **Mitigates the adverse impact of mislabelled samples**: Given the analysis in Section. 5.4.4, it is critical to reduce the adverse impact of false positives (mislabelled samples). The embeddings of mislabelled samples are often very close to the anchor, making them more likely to be easy positives. By effectively giving less weight to easy positives (mislabelled samples), thereby reducing their impact on providing misleading learning signals. Beyond sophisticated weighting, we also reduce the bias due to noisy labels, assuming access only to noisy labels. Specifically, we develop a correction for the mislabelled sample bias, leading to a new, modified loss termed debiased supervised contrastive loss (Eq. 5.19). Our approach indirectly approximates the true latent distribution, which prevents contrastive learning from being misled by incorrectly labeled samples and also reinforces the core philosophy of contrastive learning.

- **Discriminating fine detail with hard samples**: Our methodology adjusts the weighting of all samples based on their "hardness". This nuanced approach ensures that the

Table 5.7: Model accuracy measured using the acc@1 metric when trained with different loss functions on 3 popular image classification benchmarks. All models here are trained from scratch using only the indicated dataset, without pre-training.

Model	Loss	CIFAR-10	CIFAR-100	ImageNet-1K
BEiT-3	CE[203]	71.70	59.67	77.91
	SupCon[109]	88.96	60.77	82.57
	Sel-CL[128]	86.33	59.51	81.87
	TCL[94]	85.16	59.22	81.74
	Ours	**90.16**	**64.47**	**84.21**
ResNet-50	CE[203]	95.00	75.30	78.20
	SupCon[109]	96.00	76.50	78.70
	Sel-CL[128]	93.10	74.29	77.85
	TCL[94]	92.80	74.14	77.17
	Ours	**96.39**	**77.82**	**79.15**

model differentiates not only between distinctly different samples but also hones its skills on more challenging, closely related negative samples. Such an approach paves the way for a robust model that discerns in real-world scenarios where class differences might be minimal.

5.4.6 Experiments

We extensively evaluate our proposed method, SCL-RHE, on image classification in three settings: training from scratch on datasets with human-labelling errors, transfer learning using pre-trained weights (again with human-labelling errors), and pre-training on datasets with exceedingly high levels of synthetic label errors. We also conduct several ablation experiments. For all experiments, we use the official train/test splits and report the mean Top-1 test accuracy across three distinct initializations.

We employ representative models from two categories of architectures – BEiT-3/ViT base [224, 252], and ResNet-50 [81]. While new state-of-the-art models are continuously emerging (e.g. DINOv2 [175]), our focus is not on the specific choice of architecture. Instead, we aim to show that SCL-RHE is model-agnostic and enhances performance using two very different architectures.

Training from Scratch

We first evaluate our proposed SCL-RHE objective in the pre-training setting, i.e. training randomly initialized models from scratch without the use of additional data. For these experiments, we consider only human-labelling errors already present in the datasets without introducing synthetic errors. Following [109], to use the trained models for classification, we train a linear layer on top of the frozen trained models using a cross-entropy loss. We use three benchmarks:

CIFAR-10, CIFAR-100 [112], and ImageNet-1k [45]. Tab. 5.7 shows the performance of BEiT-3 and ResNet-50, with different loss functions on three popular image classification datasets. It is noteworthy that, due to the absence of pre-trained weights, BEiT-3 is identical to the ViT model [252]. We compare against training with the standard cross-entropy loss and the state-of-the-art supervised contrastive learning loss (SupCon) [109]. Additionally, we compare against two synthetic noise-mitigating contrastive learning strategies (Sel-CL [128] and TCL [94]). We see that SCL-RHE consistently improves classification accuracy over other training objectives. On Imagenet-1k, SCL-RHE leads to a 6.3% and 1.6% improvement in accuracy for BEiT-3, relative to cross-entropy and SuperCon training, respectively.

We find that SCL-RHE outperforms the existing contrastive methods Sel-CL and TCL (both designed to mitigate synthetic noise), e.g. on ImageNet-1K, SCL-RHE performs 2.4% better than Sel-CL for BEiT-3 and 2.3% better for ResNet50. We speculate that due to discarding many training pairs, Sel-CL and TCL overfit a subset of training samples, limiting their performance in the realistic setting where the rate of label noise is relatively low (e.g. 5.85% for CIFAR-100 [169]), and allowing them to be exceeded by SupCon. In contrast, SCL-RHE outperforms even SupCon, suggesting it is more applicable for real-world image training sets with low-to-moderate noise rates.

It is well known that transformer-based models underperform when training data is limited [137, 224, 293]. This is highlighted by the low performance on CIFAR-10 and CIFAR-100 with cross-entropy training. We show that SCL-RHE, and to a lesser extent SuperCon, mitigate this—relative to cross-entropy, SCL-RHE gives a 19% improvement on CIFAR-10. While BEiT-3 still fails to reach the performance of ResNet-50, the supervised contrastive approaches significantly close the gap, improving the applicability of transformer-based models in limited data scenarios. In the supplementary material, we also include an ablation study that measures the benefit of different aspects of our method. This shows that human-mislabelled samples impact the performance of supervised contrastive learning (SCL) due to their occurrence as soft positives, and that our proposed correction on both positives and negatives helps to improve performance.

Performance on corrected test sets. We next evaluate the same trained models, but using the corrected test-set labels from Northcutt *et al.* [169] (Tab. 5.8). Importantly, we can observe a relatively larger increase in performance on the corrected test sets with D-SCL—e.g. an improvement of 1.81% on ImageNet-1k. This contrasts with SuperCon [109] and cross-entropy, which show a lesser improvement of 1.17% and 0.33%, respectively. This supports the claim that D-SCL is less prone to overfitting to label noise than over SuperCon and cross-entropy. We find that Sel-CL[128] and TCL[94] generally lead to worse performance than SuperCon and do not demonstrate any performance gain when tested on the corrected labels. We speculate that, due to the relatively low mislabelling rates in these datasets (e.g. 5.85% for Imagenet), these approaches may be overly heavy-handed in combating labelling noise, diminishing the models'

Loss	Test set	CIFAR-10	CIFAR-100	ImageNet
CE[207]	Original	71.70	59.67	77.91
	Corrected	71.79 (+0.09)	59.82 (+0.15)	78.24 (+0.33)
SupCon[109]	Original	88.96	60.77	82.57
	Corrected	89.11 (+0.15)	61.49 (+0.72)	83.74 (+1.17)
Sel-CL[128]	Original	86.33	59.51	81.87
	Corrected	86.21 (-0.12)	58.62 (-0.89)	81.35 (-0.52)
TCL[94]	Original	85.16	59.22	81.74
	Corrected	84.97 (-0.19)	58.14 (-1.08)	81.28 (-0.46)
Ours	Original	90.16	64.47	84.21
	Corrected	**90.41** (+0.25)	**65.34** (+0.87)	**86.02** (+1.81)

Table 5.8: Accuracy of the BEiT-3 model using the metric acc@1 on different datasets and with various loss functions, when evaluation on both original and corrected test-set labels.

Model	FT method	CIFAR-100	CUB-200	Caltech-256	Oxford-Flowers	Oxford-Pets	iNat2017	Places365	ImageNet-1k
ViT	CE[207]	87.13	76.93	90.92	90.86	93.81	65.26	54.06	77.91
BEiT-3	CE[207]	92.96	98.00	98.53	94.94	94.49	72.31	59.81	85.40
BEiT-3	SupCon[109]	93.15	98.23	98.66	95.10	94.52	72.85	60.31	85.47
BEiT-3	Sel-CL[128]	91.48	94.52	97.19	93.71	94.51	72.43	58.36	85.21
BEiT-3	TCL[94]	90.92	93.89	97.26	93.89	94.68	72.47	59.22	85.18
BEiT-3	D-SCL (ours)	**93.81**	**98.95**	**99.41**	**95.89**	**96.41**	**76.25**	**62.53**	**86.51**

Table 5.9: Classification accuracy after fine-tuning a pretrained BEiT-3 with different loss functions, on several benchmarks.

performance as a result. D-SCL is comparatively more suitable for these lower noise rates and sees improved performance over these methods as a result.

Transfer Learning

We now assess performance when fine-tuning existing pre-trained models for specific downstream tasks. Specifically, models are initialized with publicly-available weights from pretraining on ImageNet-21k [45], and are fine-tuned on smaller datasets using our objective. We use 8 datasets: CIFAR-100 [112], CUB-200-2011 [245], Caltech-256 [72], Oxford 102 Flowers [168], Oxford-IIIT Pets [179], iNaturalist 2017 [240], Places365 [291], and ImageNet-1k [45]. We select BEiT-3 base [252] as the image encoder due to its excellent performance on ImageNet-1k. Similar to [109], our approach for fine-tuning pre-trained models with contrastive learning involves initially training the models using a contrastive learning loss, followed by training a linear layer atop the frozen trained models using cross-entropy loss.

Tab. 5.9 shows classification accuracies after fine-tuning with different methods. We see that D-SCL gives the best classification accuracy across all datasets, with particularly large improvements on iNat2017 (+3.4%) and state-of-the-art performance on Places365 (+2.2%) when compared to fine-tuning with SupCon [109]. Similar to the pre-training setting, the noise-robust objectives Sel-CL and TCL exhibit inferior performance compared to fine-tuning with cross-

Loss	CIFAR-10			CIFAR-100		Time
Noise level	*Original*	*20%*	*40%*	*Original*	*40%*	
CE[207]	91.84	82.02	76.86	73.74	45.05	**18.3**
SupCon[109]	94.08	89.13	79.57	74.58	51.33	20.2
Sel-CL[128]	91.42	94.45	**93.22**	72.10	74.24	29.1
TCL[94]	90.78	93.96	93.13	72.18	**74.62**	32.7
Ours	**95.91**	**94.71**	92.59	**77.62**	74.08	20.4

Table 5.10: Performance of ResNet-18 trained at different synthetic-noise levels. Time (Min/epoch) means the training time on a Nvidia A6000.

entropy or D-SCL across all 8 datasets.

Robustness to Synthetic Noisy Labels

Whereas standard datasets exhibit only a moderate level of label noise (e.g. 5–10%), we now examine performance at much higher noise rates (>18%) than is typical in benchmarks. As in Section. 5.4.6, we use BEiT-3 trained from scratch without additional data. Specifically, we use the artificially noisy datasets CIFAR-10N [254] and CIFAR-100N [254]. Crucially, the label noise in these datasets is predominantly *not* due to naturally-arising human errors, but rather to synthetic mislabeling. This breaks our hypothesis (and foundational principle of D-SCL) that mislabelled samples usually have high visual similarity to their assigned class and occur naturally due to human errors.

Tab. 5.10 shows that D-SCL outperforms cross-entropy and SuperCon on noisy and noise-free variants of these datasets. Again, we compare against Sel-CL and TCL and find that D-SCL's performance diminishes relative to these methods for high levels of label noise (e.g. 40%). However, we show that D-SCL remains competitive with these methods at noise levels up to 18% and significantly outperforms them when no noise is present. As such, we argue that D-SCL is more applicable and well suited to more realistic scenarios where the label noise rate is relatively low.

5.4.7 Discussion and Summary

Limitations. Although SCL-RHE exceeds the state-of-the-art, it still has certain limitations. First, while existing research has estimated mislabelling rates for various datasets [169], determining this for new datasets remains a challenge. This issue could be effectively managed by adopting the typical average error rate of 3.3%, as reported in [169], as a baseline for hyper-parameter tuning to identify an optimal value. Moreover, in our experiments, we observed that SCL-RHE's performance exhibits low sensitivity to the estimated mislabelling rates. Second, although our SCL-RHE outperforms existing methods even for mitigating synthetic errors up to

20% noise rates, it does not surpass Sel-CL [128] and TCL [94] in scenarios involving extremely high synthetic error rates of 40%. This is because our model is tailored to mitigate human-labelling errors in datasets with error characteristics typical of real-world scenarios—i.e. where mislabellings occur naturally due to human error when classes are genuinely similar or ambiguous.

Summary. In this section, we investigated the extent and manner in which human-labelling errors impact supervised contrastive learning (SCL), and demonstrated these impacts diverge from those on regular supervised learning. Based on this, we introduced a novel SCL objective that is robust to human errors, SCL-RHE, specifically designed to mitigate the influence of real human-labelling errors (instead of synthetic noise addressed in previous works). Our empirical results reveal that SCL-RHE consistently outperforms traditional cross-entropy methods, the previous state-of-the-art SCL objectives, and noise-mitigating approaches designed for synthetic noise, both when training from scratch and in transfer learning. In addition to its superior performance, a key advantage of SCL-RHE is its efficiency—unlike previous methods that mitigate synthetic label noise, it incurs no extra overhead during training.

5.5 Conclusion

In this chapter, we introduced three methods that utilize contrastive learning to address research question 1 in Section 4.3, specifically examining how contrastive learning methods impact modality alignment. These methods aim to resolve the shallow intra-modal and inter-modal alignment issues in multimodal learning. Visualizations of the embeddings produced by our methods show superior class separation and reduced overlap between positive and negative samples compared to other approaches, indicating improved discriminative capabilities. Additionally, we consistently observed enhanced performance across a wide range of datasets when applying our methods, confirming that the improved embeddings lead to better outcomes in downstream tasks. Overall, our proposed methods enhance the generalization and transferability of Vision Transformers, produce high-quality visual embeddings, and improve intra-modal alignment from various perspectives, thereby supporting the thesis statement that enhancing alignment significantly improves multimodal learning performance. In particular:

- In Section 5.2, we introduce *LaCViT*, a label-aware contrastive fine-tuning framework that significantly improves the Top-1 accuracy of vision transformers across multiple benchmarks. *LaCViT* provides a versatile and comprehensive strategy that greatly enhances the efficacy of transformers for image classification. Our thorough empirical evaluations confirm *LaCViT*'s effectiveness and position it as a viable alternative to the traditional cross-entropy method for fine-tuning pre-trained image classification models. Extensive experimentation across eight image classification datasets shows that *LaCViT* significantly outperforms baseline models, such as a 10.78% increase in Top-1 Accuracy for the *LaCViT*-trained MAE on the CUB-200-2011 dataset [115].

- In Section 5.3, we present CLCE, a method that integrates label-aware contrastive learning with hard negative mining and cross-entropy (CE) to overcome the limitations of CE and existing contrastive learning techniques. Our empirical data show that CLCE surpasses both traditional CE and earlier contrastive learning methods in both few-shot and transfer learning contexts. Importantly, CLCE is particularly suitable for researchers and developers with access to only commodity GPU hardware, as it achieves effective performance with smaller batch sizes that fit on less powerful GPUs. Our extensive experiments show the state-of-the-art performance of our CLCE in Few-Shot Learning and Transfer Learning settings: CLCE significantly surpasses CE by an average of 2.74% in Top-1 accuracy across four few-shot learning datasets when using the BEiT-3 base model, with large gains observed in 1-shot learning scenarios. Additionally, in transfer learning settings, CLCE consistently outperforms other state-of-the-art methods across eight image datasets, setting a new state-of-the-art result for base models (88 million parameters) on ImageNet-1k.

- In Section 5.4, we explore how human labeling errors affect supervised contrastive learning (SCL) differently than they do traditional supervised learning. In response, we develop a new SCL objective, SCL-RHE, which is resistant to real human labeling errors. Our empirical findings indicate that SCL-RHE consistently outperforms traditional cross-entropy approaches, previous SCL objectives, and noise-correcting methods tailored for synthetic noise, in both initial training and transfer learning scenarios. SCL-RHE also stands out for its efficiency—it does not require additional training overhead, unlike methods designed to correct synthetic label noise. Our experiment result shows that SCL-RHE gives the best classification accuracy across all evaluated datasets, with particularly large improvements on iNat2017 (+3.4%) and state-of-the-art performance on Places365 (+2.2%) when compared to fine-tuning with SupCon.

In the next chapter, we outline our efficient transfer learning methods that significantly accelerate the transfer process by employing adapter techniques. These techniques facilitate the practical application of our proposed methods across various domains, while also leveraging the contrastive learning methods discussed in this chapter.

Chapter 6

Efficient Multimodal Large Language Model Learning

6.1 Introduction

In this chapter, we introduce the efficiency component illustrated in Figure 4.1, which aims to address the efficiency issues that limit the practical applications of the proposed multimodal learning framework, as stated in the thesis statement (Section 1.2). Simultaneously, we aim to further mitigate the shallow intra-modal and inter-modal alignment problems to enhance effectiveness through dedicated model design and the contrastive learning methods proposed in Section 5.

Recall from Section 1.1 that we identified the rising computation costs for transfer learning across various tasks, driven by the growth of large language models and their exponentially increasing sizes. Indeed, recent advancements in Multimodal Large Language Models (MLLMs), such as BLIP2 [123] and BEiT-3 [252], have demonstrated state-of-the-art performance in multimodal tasks, exemplified by their capabilities in Visual Question Answering. However, the adaptation of these MLLMs to specialized downstream tasks remains a substantial challenge, particularly for image-text retrieval, a common use-case in multimodal learning. Traditional full fine-tuning requires isolated, exhaustive retraining for each new task, demanding intensive computational resources and thus limiting practical applications. For instance, training BLIP2-Giant on an Nvidia A100 GPU takes 144 days [123].

Given the challenge of fine-tuning MLLMs, there is a growing need to develop efficient adaptation methods for MLLMs [88, 228]. While progress has been made in unimodal domains using adapter modules, these methods remain largely underexplored in multimodal contexts, particularly for image-text retrieval. Furthermore, existing adaptation methods for MLLMs [28, 228, 279] focus on information extraction from downstream datasets but neglect the critical need for inter-modal alignment. The goal of inter-modal alignment is to bring different modalities into a common feature space where they can be effectively compared, combined, or related. With

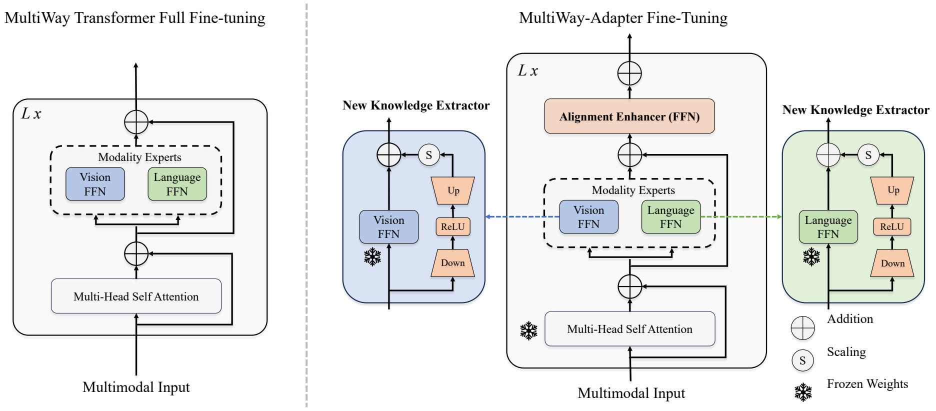


Figure 6.1: **Comparison of a MultiWay Transformer and our MultiWay-Adapter fine-tuning.** MultiWay-Adapter uses a dual-component design, including New Knowledge Extractor and Alignment Enhancer. We replace the original FFN with New Knowledge Extractor: frozen branch (left) and the trainable bottleneck module (right). Moreover, we add a Alignment Enhancer upon the original FFN to enhance the inter-modal alignment.

shallow alignment, the model would fail to capture the complex inter-relations between different modalities, thereby impacting its effectiveness in multi-modal tasks [14, 147, 225].

To address the issue of shallow inter-modal alignment while preserving the efficiency advantages of adapter approaches, we introduce the MultiWay-Adapter (MWA), a lightweight yet effective framework designed explicitly for MLLMs adaptation. Additional components of MWA are small in size but bring a significant performance boost in transfer learning with minimal fine-tuning cost, which are compatible with components proposed in previous chapters in this thesis. Our key contributions in this chapter include:

- We propose MWA, a framework incorporating a dual-component approach: the New Knowledge Extractor and the Modality Enhancer. MWA extracts new knowledge from downstream datasets while ensuring deep inter-modal alignment, which is crucial for superior performance in vision-language tasks. To our knowledge, this is the first work to address the issue of shallow inter-modal alignment in adapter approaches for MLLMs, directly supporting our thesis statement on improving alignment and performance.

- Through comprehensive experiments, we demonstrate that MWA achieves superior zero-shot performance on the Flickr30k dataset by tuning only an additional 2.58% of parameters in the BEiT-3 Large model. This approach saves up to 57% in fine-tuning time compared to full-model fine-tuning, without statistically significant decreases in performance in other settings. This efficiency aligns with our thesis statement on enhancing multimodal learning frameworks while reducing resource requirements.

- Experimental results show the robustness of MWA as parameters scale up, making it well-suited for MLLMs that are continually increasing in size. This scalability supports our thesis claim of developing efficient and adaptable multimodal learning methods.

- Our ablation study confirms the effectiveness of both MWA components, validating our design choices and further substantiating the thesis statement that targeted enhancements in multimodal learning frameworks lead to significant performance improvements.
- The original material in this section has been accepted for presentation at the 2024 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), a conference with an h5-index of 123.

6.2 MultiWay-Adapter: Adapting Multimodal Large Language Models for scalable image-text retrieval

We introduce MultiWay-Adapter (MWA), designed for the efficient transfer of Multimodal Large Language Models (MLLM) to downstream tasks. Although the primary focus of this paper is on image-text retrieval tasks, the potential applicability of the MWA is broader, such as video text retrieval and image captioning.

6.2.1 Preliminaries

The overall framework is constructed on the basis of a popular architecture of MLLM, which utilizes a MultiWay Transformer design [252]. As depicted on the left of Figure 6.1, each MultiWay Transformer block comprises a shared self-attention module and a pool of feed-forward networks (i.e., modality experts) tailored for different modalities. This design is similar to the dual-backbones architecture of multimodal models, e.g., one encoder for vision input and another encoder for language input, yet differs by sharing the weights within each self-attention module. This design choice reduces the parameter count and enhancing inter-modal alignment—an essential quality for high-performance multimodal tasks [252].

6.2.2 MultiWay-Adapter

Overall Architecture.

Our proposed MWA uses a dual-component approach: the New Knowledge Extractor and the Alignment Enhancer, as illustrated on the right of Figure 6.1.

New Knowledge Extractor.

The New Knowledge Extractor is designed for extracting new knowledge from the target downstream tasks. In contrast to the conventional full fine-tuning of MultiWay Transformers, we replace both feed-forward networks (FFNs) in the transformer block with a New Knowledge

					MSCOCO (5k test set)		Flickr30k (1k test set)	
Model	FT-Way	Tunable params (M)	GPU Mem (GB)	Time (Min)	IR@1	TR@1	IR@1	TR@1
ALBEF [125]	Full Fine-tune	196	N/A	N/A	60.7	77.6	85.6	95.9
ALIGN [100]	Full Fine-tune	825	N/A	N/A	59.9	77.0	84.9	95.3
BEiT-3-Base	Full Fine-tune	222 (100%)	37GB	225	61.4	79.0	86.2	96.3
BEiT-3-Large	Full Fine-tune	675 (304%)	45GB	353	63.4 (+2.0)	82.1 (+3.1)	88.1 (+1.9)	97.2 (+0.9)
BEiT-3-Base	MultiWay-Adapter	7.13 (**3.21%**)	**30GB**	**130**	60.7 (-0.7)	78.3 (-0.7)	85.4 (-0.8)	95.4 (-0.9)
BEiT-3-Large	MultiWay-Adapter	17.40 (**2.58%**)	**36GB**	**194**	63.3 (+1.9)	82.1 (+3.1)	88.0 (+1.8)	97.1 (+0.8)

Table 6.1: **Comparative Analysis of Full Fine-Tuning and the MultiWay-Adapter**: The table shows Top-1 recall metrics on COCO and Flickr30k datasets, presented as both absolute values and relative gaps to the BEiT-3 Base full fine-tuning Model. Metrics for Text-to-Image Retrieval (IR) and Image-to-Text Retrieval (TR) are provided. GPU memory usage and training time are also included. Training time is measured using a single NVIDIA A6000 GPU with 48GB memory for one epoch.

Extractor. This extractor comprises two branches: the left branch, identical to the original network, and an additional right branch introduced for task-specific fine-tuning. The latter utilizes a bottleneck structure to limit the number of parameters and includes a down-projection layer and an up-projection layer. Formally, for a specific input feature $x_i{}'$, the right branch of the New Knowledge Extractor produces the adapted features, $\tilde{x}_i$, as:

$$\tilde{x}_i = \mathrm{ReLU}(\mathrm{LN}(x_i{}') \cdot \mathbf{W}_{\mathrm{down}}) \cdot \mathbf{W}_{\mathrm{up}} \tag{6.1}$$

Here, $\mathbf{W}_{\mathrm{down}} \in \mathbb{R}^{d \times \check{d}}$ and $\mathbf{W}_{\mathrm{up}} \in \mathbb{R}^{\check{d} \times d}$ denote the down-projection and up-projection layers, respectively. $\check{d}$ is the bottleneck middle dimension and satisfies $\check{d} \ll d$. LN denotes LayerNorm. This bottleneck module is connected to the original FFN (left branch) through a residual connection via a scale factor α. Then, these features, $x_i{}'$ and $\tilde{x}_i$, are fused with the original one, x_i, through a residual connection:

$$x_i = FFN(LN(x_i{}')) + \alpha \cdot \tilde{x}_i + x_i' \tag{6.2}$$

Alignment Enhancer.

After extracting new knowledge from the target downstream task, to maintain and improve inter-modal alignment, an Alignment Enhancer module is added atop the pool of feed-forward networks. This module mimics the architecture of the New Knowledge Extractor but uses a larger middle dimension to facilitate better feature fusion and alignment.

During the fine-tuning phase, only the parameters of these newly added modules are optimized, while the rest of the model is frozen (as indicated by the frozen sign in Figure 6.1). This strategy makes MWA a plug-and-play module, applicable to other MLLM, such as CLIP [189], VLMo [15], and ALIGN [100].

6.2.3 Experiments

Setup

We conducted experiments on two state-of-the-art MLLMs, BEiT-3 Base and BEiT-3 Large, across two widely-used image-text retrieval datasets: MSCOCO [133] and Flickr30K [185]. We use the 5k test set of MSCOCO and 1k test set of Flickr30k to report metrics, in accordance with previous studies [133, 185]. We initialized the backbone, excluding our additional modules, with pre-trained weights, which were frozen during the fine-tuning process when employing MultiWay-Adapter. For fine-tuning, the batch size is 512 for the Large model and 1024 for the Base model, over 20 epochs with an initial learning rate of 0.001. Middle dimensions for the New Knowledge Extractor and the Alignment Enhancer were set to 64 and 128, respectively. All the code used in our experiments can be found in `https://github.com/longkukuhi/MultiWay-Adapter`.

Experimental Results

The objective of this experiment is to assess the efficiency and efficacy of our MWA framework in comparison to traditional full fine-tuning methods. We compared our MWA approach with full fine-tuning in two distinct settings: fine-tuning performance and zero-shot performance.

Fine-Tuning Performance

As shown in Table 6.1, our MWA method demonstrates superior computational efficiency. Specifically, it utilizes a mere 3.21% and 2.58% of the trainable parameters for the Base and Large variants of BEiT-3, respectively, in contrast to conventional full fine-tuning. This leads to a substantial reduction in GPU memory consumption—by 7GB and 9GB for the Base and Large variants, respectively. Furthermore, MWA significantly reduces the time required for fine-tuning. For instance, fine-tuning MWA with the BEiT-3 Base model is reduced by 57% compared to full fine-tuning.

Regarding effectiveness, the performance decrement when utilizing MWA is statistically insignificant for both the Base and Large BEiT-3 variants, with deviations falling within a margin of less than 1%. Synthesizing these efficiency and effectiveness attributes demonstrates that MWA, when applied to the BEiT-3 Large model, consumes merely 86% of the time required for full fine-tuning of the BEiT-3 Base model, yet surpasses its performance. This suggests that MWA enables enhanced performance with reduced computational time, particularly for larger models. Additionally, as the model size increases, the performance disparity between MWA and full fine-tuning diminishes, indicating a positive correlation between MWA's effectiveness and model size.

		Flickr30k	
Model	FT-Way	IR@1	TR@1
BEiT-3-Large	Full fine-tune	85.99	95.48
BEiT-3-Large	MultiWay-Adapter	86.26	95.51

Table 6.2: **Zero-shot performance on Flickr30k**.

Zero-Shot Performance

To evaluate the transfer capabilities of MWA and full fine-tuned methods, we conducted experiments in a zero-shot setting. In this setting, the model is evaluated on Flickr30k (1k test set), with which it has no prior knowledge of, thereby necessitating reliance on intrinsically learned knowledge to simulate the handling of previously unseen samples. These models were initially fine-tuned on the MSCOCO dataset. As shown in Table 6.2, MWA surpasses the performance of full fine-tuning when employed with the BEiT-3 Large model. We hypothesize that this enhancement is attributable to the preservation of generalizable knowledge in the frozen weights, knowledge potentially lost during the full fine-tuning process. This retained knowledge augments the model's ability to adeptly manage unseen instances. Thus, MWA not only match the performance of full fine-tuning method but also distinguishes itself in terms of resource efficiency and transferability.

In summary, the experimental results demonstrate that MWA serves as an effective and resource-efficient fine-tuning method for MLLMs, especially when computational resources are constrained.

6.2.4 Analysis

Scaling Tunable Parameters Up

The primary aim of this section is to investigate the impact of varying the number of tunable parameters on performance and to identify the optimal value for additional parameters. The "mid-dimension" of the New Knowledge Extractor largely controls the number of tunable parameters. We conducted an empirical evaluation across a range of mid dimensions {0, 1, 16, 32, 64, 128} on the MSCOCO dataset using the BEiT-3 Base model. The results are summarized in Figure 6.2. The data reveals a noticeable increase in performance as the dimension grows, plateauing at 64. Specifically, we observed a peak performance gain of 9.45%, in text to image retrieval when increasing the dimension from 1 to 64. This indicates that increasing the number of parameters in the adapter does not guarantee performance improvement. When the dimension is set to zero, it represents the zero-shot performance of the BEiT-3 Base model without MWA. Notably, MWA delivers superior performance compared to the zero-shot performance of

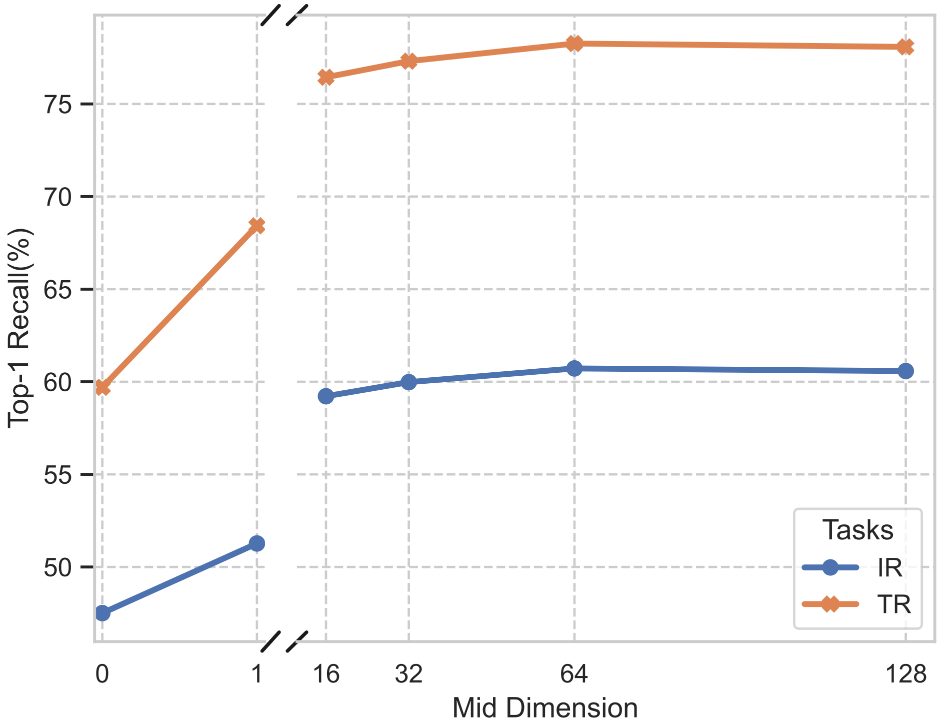


Figure 6.2: **Evaluation of different sizes of mid-dimension New Knowledge Extractor on MSCOCO**.

the BEiT-3 Base model, even when the mid-dimension is as low as one. Furthermore, performance variability is relatively small when increasing the dimension from 16 to 64, indicating that MWA is stable in tuning and not sensitive to changes in size.

Ablation on MultiWay Adapter's Components

In this section, our focus is to quantify the individual contributions of our two newly introduced components: the New Knowledge Extractor and the Alignment Enhancer. An ablation study was performed on the MSCOCO dataset using the BEiT-3 Base model. The performance metrics for each component, both in isolation or in combination, are detailed in Table 6.3. Our findings demonstrate that omitting either component leads to a significant decline in performance, approximately 3%, for image to text retrieval and around 4%, for text to image retrieval. Importantly, the Alignment Enhancer, a novel element distinct from previous Adapter methods, validates its critical role in maintaining deep alignment between modalities through observed performance gains. In summary, both components not only significantly contribute to the overall performance but also complement each other effectively.

6.2.5 Conclusion

In this chapter, we introduce the MultiWay-Adapter (MWA) to address research question 2 raised in Section 4.3, providing an effective method for the efficient adaptation of Multimodal Large

			MSCOCO	
Model	KE	AE	IR@1	TR@1
BEiT-Base			61.40	79.00
BEiT-Base	✓		57.32	73.92
BEiT-Base		✓	57.88	74.61
BEiT-Base	✓	✓	60.72	78.26

Table 6.3: **Ablation study of two modules of MultiWay-Adapter**. KE refers to the New Knowledge Extractor and AE refers to the Alignment Enhancer.

Language Models (MLLM) to downstream tasks. Addressing the issue of shallow intra-modal and inter-modal alignment in existing methods, MWA employs a dual-component approach, utilizing both the New Knowledge Extractor and the Alignment Enhancer. This strategy enables MWA to extract novel information from downstream datasets while securing deep inter-modal alignment. Our empirical findings reveal that with the addition of only 2.58% extra parameters, there is no statistically significant decline in performance across all tested settings while reducing fine-tuning time by up to 57%.

The motivation behind developing MWA lies in its ability to enhance the efficiency and alignment of multimodal learning frameworks, which is crucial for improving overall performance and applicability in real-world scenarios. By effectively addressing the core issues stated in the thesis statement (see 1.2), MWA becomes an integral component of our proposed multimodal learning framework, MCA.

To comprehensively validate our proposed framework, it is essential to evaluate its performance across diverse and challenging domains. Therefore, the following chapters will investigate the applications of MCA in four distinct areas: crisis response, cross-modal retrieval, robotic vision, and recommendation systems. These evaluations will demonstrate the versatility, robustness, and practical effectiveness of MCA, providing a thorough assessment of its impact and potential across various real-world tasks.

Chapter 7

The Applications of Proposed Multimodal Framework in Crisis Response

7.1 Introduction

In this chapter, we aim to evaluate whether our proposed multimodal learning framework, MCA, improves performance in the context of crisis response by addressing the intra-modal and inter-modal alignment issues outlined in the thesis statement (see Section 1.2). We expect to see improvements in performance metrics such as accuracy and F1 score, thereby demonstrating the effectiveness of our MCA framework.

Social media platforms, like Twitter, are increasingly used by billions of people internationally to share information. As such, these platforms contain vast volumes of real-time multimedia content about the world, which could be invaluable for a range of tasks such as incident tracking, damage estimation during disasters, insurance risk estimation, and more. By mining this real-time data, there are substantial economic benefits, as well as opportunities to save lives.

In Section 7.2, we explore the feasibility of utilizing automated methods to classify social media information by analysing multimodal data which encompasses both visual and linguistic elements. Indeed, as people post everything they experience on social media, large volumes of valuable multimedia content are being recorded online, which can be analysed to help for a range of tasks. However, the majority of prior works in this space focus on using machine learning to categorize single-modality content (e.g. text of the posts, or images shared), with few works jointly utilizing multiple modalities. Hence, in Section 7.2, we examine to what extent integrating multiple modalities is important for crisis content categorization. In particular, we design a pipeline for multi-modal learning that fuses textual and visual inputs, leverages both, and then classifies that content based on the specified task. Through evaluation using the CrisisMMD dataset, we demonstrate that effective automatic labelling for this task is possible, with an average of 88.31% F1 performance across two significant tasks (relevance and humanitarian category classification). while also analysing cases that unimodal models and multi-modal models success and fail.

7.2 Is Multi-Modal Data Key for Crisis Content Categorization on Social Media?

7.2.1 Introduction

We combine both types of data to perform classification tasks on social media content in a multimodal manner in this section.

Indeed, previous works have primarily focused on analysing the text within each post [115, 224, 236], ignoring potentially valuable image data provided along-side the tweets. One of the more significant unresolved challenges is multi-modal understanding, i.e. for some content we need to concurrently analyse both modalities (text and image) before we can fully understand and act upon it. For example, a user could post "need help" with an image of a damaged building, where the text defines the request and the image provides the location. On the other hand, with recent advancements in deep learning for computer vision [81] and language modelling [46], multi-modal learning[63] provides a promising new direction to push the boundaries of effectiveness in this domain. Using a large-scale visual recognition model trained from ImageNet, we can extract extra information in attached images and combine this with textual evidence to enhance crisis content analytic tasks, where this type of multi-modal learning pipeline can be applied to tweets that only have text, or images, or the combination of both.

To address this challenge and improve the performance of crisis response, we propose a new multi-modal framework to classify crisis-related tweets through analysis of textual and visual content. Following the task definition of the CrisisMMD dataset [2], we present a novel approach to label tweets on two major tasks automatically:

- **Informativeness**: Whether the social media post is useful for providing humanitarian aid during emergencies.
- **Humanitarian Information Categories**: Identifying the type of emergency, including affected individuals, rescue volunteering or donation effort, infrastructure and utility damage.

The contributions of this work are three-fold and directly support our hypothesis and thesis statement by enhancing the performance and applicability of multimodal learning frameworks in crisis response.

1. We propose a general multimodal framework capable of classifying tweets with multimodal data in the crisis domain. This supports our thesis statement by demonstrating the effectiveness of integrating textual and visual data to improve crisis response tools.

2. We analyze various classification layer designs building upon pre-trained ResNet and BERT models, addressing practical questions regarding the training of multimodal models. This analysis validates our hypothesis that tailored designs and training strategies can significantly enhance model performance, supporting the thesis claim of improved alignment and efficiency.

3. We discuss notable insights gained from analyzing the developed multimodal models, particularly regarding their success and failure modes. These insights provide valuable guidance for further optimizing multimodal learning frameworks, reinforcing our thesis statement by highlighting the practical benefits and understanding achieved through our proposed methodologies.

Based on experimentation over the CrisisMMD dataset [2], we show that effective automated tooling to aid in the filtering of crisis-related tweets for emergency personnel is possible with around 90% F1 performance on the informativeness task and 87% F1 on the categorization task. Comparing uni-modal models, we demonstrate that text-only models are more effective than images models when tuned (e.g. 86% vs. 84% F1 on informativeness). Beyond this, we introduce a novel multi-modal framework and demonstrate that it outperforms both strong uni-modal and multi-modal baselines, as well as an existing multi-modal approach by up to 5% absolute F1 on the informativeness task and 8% absolute F1 on the categorization task, demonstrating that considering multi-model data is key for these tasks.

The remainder of this section is structured as follows. In the next section, we describe training methodology employed and discuss factors that might affect the performance of our pipeline and details of our methodology to build a multi-modal learning pipeline for the crisis response, followed by a structured overview of our experiment setup, where we provide more technical details about the dataset. Finally, we report our results, analysis and summary.

7.2.2 Methodology

In this section, to evaluate whether integrating both text and images is actually important for crisis content categorization, we first define strong uni-modal baselines. Notably, these uni-modal baselines need to be as effective as possible, since if we use older components like past works [170, 295] did, then any gains observed from multi-modal combinations might be similarly achieved by making the baselines stronger. For this reason, we first start with the most effective uni-modal pre-trained models from the literature, tune them for the target task, and then optimise the training hyperparameters. We discuss the implementation of these models below:

Uni-Modal: Text: When implementing a text categorization model, there are two main decisions that need to me made: 1) how to embed the text; and 2) how to train the classification layer. For the embedding, we experiment with two approaches:

- **Word2vec**: Although not the focus of this work, we include a shallow word-embedding approach as a point of comparison. In particular, we report the performance of well known Word2Vec model trained using the Continuous Bag-of-Words (CBOW) approach [160]. In particular, we use the same word-embedding model as [170].

- **BERT**: As discussed earlier, state-of-the-art text embeddings use pre-trained transformer-based models, hence we do the same using BERT [46]. There are two frequently used versions of the BERT, BERT-Base uncased and BERT-Large uncased. We use BERT-base uncased (referred to as BERT-base) here due to the very high memory overheads of BERT-Large uncased. As input to BERT, tweets are first subject to stopword removal using SPACY[1], followed by the default word-piece tokenizer [194]. The pre-trained BERT model expects a fixed length input. To achieve this, we perform zero-padding for any tweets with fewer than 100 (word-piece) tokens. Following best practices, we fine-tuned the pre-trained BERT model on the training/validation components of the CrisisMMD dataset (AdamW optimiser and ReLU activation function) with a softmax output layer.

Meanwhile, for the classification model we experiment with three variants of neural layer on-top of the text embedding, summarized below. For the remainder of this section, we will use a short-hand notation to refer to the construction of the classification model/layers used, as follows. **Dense(n)** refers to *n* dense fully connected layers, while **Conv(n)** refers to *n* convolutional layers. **Norm** denotes a batch normalization layer, while **DO** denotes a drop-out layer. **A/B(n)** represents a grouping of *n* **B** layers with *n-1* **A** layers in-between. The **|** symbol is used to connect layers. Later when we discuss multi-modal models, **Concat** is used to denote the concatenation of two dense layers, one from each modality.

- **DO|Dense(1)**: A single drop-out layer (rate of 0.1) followed by a single dense neural layer. In effect, the single dense layer acts similarly to a linear regression model.

- **DO/Dense(3)**: Three fully connected neural layers with decreasing sizes of 200, 100 and 50. Two dropout layers are applied between these three layers to prevent over-fitting, with dropout rates of 0.2 and 0.02, respectively. This provides a more expressive classification layer for the down-stream task.

- **Norm/Conv(3)|Dense**: As a point of comparison, we include the same configuration of head as [170]. This approach uses a CNN-based classifier, comprised of three convolutional layers with normalization layers in between, followed by a dense fully connected layer. The convolutional layers of increasing size [100,150,200] and use kernel sizes/pooling lengths [2,3,4].

[1] `https://spacy.io/`

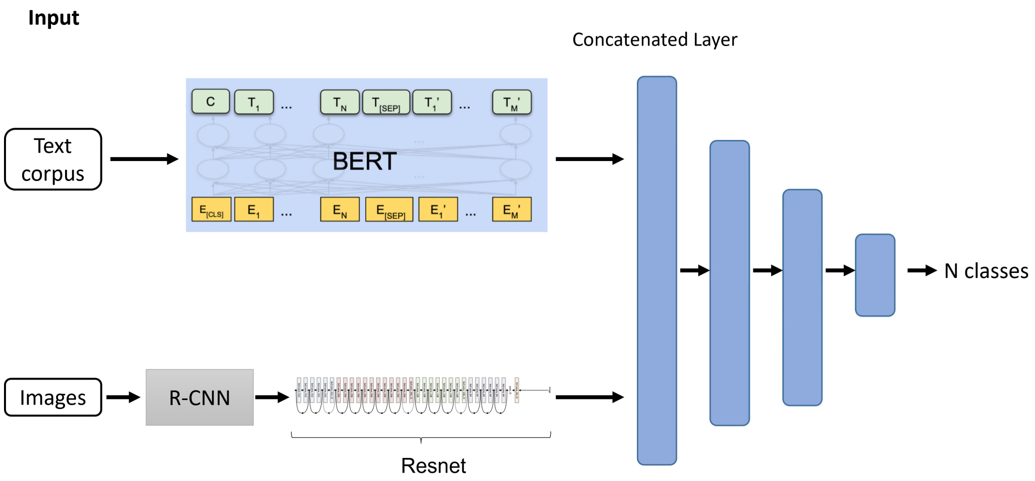


Figure 7.1: Details of models architecture

All three variants are trained on the training/validation components of the CrisisMMD dataset using either the Adam or AdamW optimiser (reported in the result table), a batch size of 32 and using the ReLU activation function.

Uni-Modal: Images: For training uni-modal image categorizers, the primarily influencing factor is the pre-trained model that we use to generate the image embeddings. For our later experiments, we compare three different pre-trained deep neural image models:

- **VGG16**: A CNNs based network proposed by [214] which has 16 convolutional layers and performs well on a wide range of tasks. It is pretrained on ImageNet dataset [204].
- **ResNet50**: A variant of ResNet with 50 layers. It is pretrained on ImageNet dataset [204].
- **ResNet152**: A variant of ResNet with 152 layers. It is pretrained on ImageNet dataset [204].

As these image models are designed for classification, rather than adding a new layer on top of the existing model, we instead replace their final dense classification layer with a new one for our target task (using our notation from earlier, this would simply be Dense(1)). We trained this layer on the training/validation components of the CrisisMMD dataset using the Adam optimiser, a batch size of 16 and the ReLU activation function. For ResNet152, we also experimented with larger batch sizes [16,32,64,128], that we report later.

Multi-Modal: Text+Images: Having described the different uni-modal model baselines that we use later for comparison, we next describe how we integrate these models to form an effective multi-modal model. As discussed in the related work, there are two broad strategies for multi-modal construction: early interaction; and late interaction. Since we wish to combine existing single-modality models that have different architectures (a transformer for text and a CNN for images), our only option here is to follow a late interaction strategy, as illustrated in Figure 7.1. In particular, we experiment with two classification layer configurations on top of the embedding components of the uni-modal models:

- **Concat|Norm|DO|DO/Dense(2)**: We add one extra dense layers with dimension 1200 for each of the two modalities, that are concatenated. These features are then passed through a batch normalisation layer, followed by two more dense layers (sizes [500,100] with dropout layers in between (dropout rates [0.4,0.2,0.02])).

- **Concat|Conv(3)|Dense**: In a similar way to the text uni-modal, we also compare against the fusion classification layer used by [170]. Like for the text-only variant, this approach uses a CNN-based classifier. However, this is only comprised of two convolutional layers (size [500,100], followed by a dense fully connected layer).

As for the uni-modal models, we trained on the training/validation components of the CrisisMMD dataset using the Adam optimiser and the ReLU activation function.

7.2.3 Experimental setup

Dataset: Currently the most widely used multi-modal crisis dataset is CrisisMMD [2]. This dataset contains provides tweets and human annotated labels for two crisis informatics tasks:

- **Informative vs. Not Informative**: A given tweet text or image whether is useful for humanitarian aid purposes, defined as useful for providing assistance to people in need.

- **Humanitarian Categories**: Given an image, or tweet, or both, categorize it into one of the following categories: infrastructure and utility damage; vehicle damage; rescue, volunteering, or donation efforts; injured or dead people; missing or found people; other relevant information; not humanitarian related.

This dataset provides a total of 16,058 tweet texts and 18,082 tweet images labelled for these two tasks. However, just because a text and image pair comes from the same tweet, this does not mean that they have the same label. Hence following prior work [170] we only use tweets where the label for the text and image in a tweet agree. For the purposes of training and testing our models, we use the same train/validation/test split as [170].

Metrics: We evaluate the performance of both the uni-modal and multi-modal models produced via the following classical classification metrics: accuracy, precision, recall and weighted F1-score. Note that all metrics are reported on the same test set of CrisisMMD. For Humanitarian Categorization, the prevalence of categories uneven in the test set. As such, we focus more on F1 as a primary metric as opposed to accuracy (which is biased towards the most represented categories).

Baselines: Our overall goal in this section is to determine to what extent we need to consider multi-modal evidence when performing crisis content categorization. Hence, our primary comparison will be between the best uni-modal models that we can produce and the associated multi-modal models. However, as a recent paper that tackled the same task, we compare against

			Informativeness Task					Humanitarian Categorization Task				
Text Embedding	Classification layer	Optimizer	Acc	precision	recall	F1	Time	Acc	precision	recall	F1	Time
Word2Vec	DO|Dense(1)	adam	0.6565	0.5294	0.6565	0.5441	00:01:36	0.5246	0.278	0.5246	0.3634	00:01:32
Word2Vec	DO/Dense(3)	adam	0.5821	0.5582	0.5821	0.5676	00:02:35	0.4293	0.3644	0.4293	0.3813	00:02:23
Word2Vec	Norm/Conv(3)|Dense	adam	0.8080	0.8100	0.8100	0.8090	00:15:27	0.7040	0.7000	0.7000	0.6770	00:25:00
BERT-Base	DO|Dense(1)	adam	0.7927	0.8077	0.7927	0.7694	00:18:54	0.8063	0.8132	0.8063	0.8081	00:12:31
BERT-Base	DO/Dense(3)	adam	0.8201	0.829	0.8201	0.8058	00:18:42	0.8126	0.8188	0.8126	0.8137	00:12:36
BERT-Base	Norm/Conv(3)|Dense	adam	0.8627	**0.8644**	0.8627	**0.8644**	01:55:21	0.8084	0.8129	0.8084	0.8067	01:53:36
BERT-Base	DO|Dense(1)	adamw	0.8514	0.8503	0.8514	0.8507	00:13:08	0.8168	0.818	0.8168	0.8168	00:13:21
BERT-Base	DO/Dense(3)	adamw	**0.8651**	0.8638	**0.8651**	0.8620	00:19:42	0.8115	0.8177	0.8115	0.8121	00:13:32
BERT-Base	Norm/Conv(3)|Dense	adamw	0.8631	0.8622	0.8631	0.8625	01:55:01	**0.8304**	**0.8345**	**0.8304**	**0.8318**	01:55:01

Table 7.1: Crisis Content Categorization of Uni-Modal: **Text** models.

[170] as a multi-modal baseline as well. For reference, this uses Word2Vec for test embedding and VGG16 for Image embedding, followed by a **Concat|Norm/Conv(3)|Dense** classification head.

7.2.4 Experimental results

In this section we report the results comparing the performances of both the uni-modal text, uni-modal image and multi-modal (text+image) models produced for crisis content categorization. In particular, we divide this section into two main components: 1) Finding Effective Uni-Modal Models, where we optimise the uni-modal models to form strong baselines to compare to; and 2) Uni-Modal vs. Multi-Modal Comparison, where we contrast the performance of the uni-modal and multi-modal models.

Finding Effective Uni-Modal Models Based on the variables previously described in the Methodology section, we train a range of uni-modal models for both CrisisMMD tasks. Primarily, we are looking to find the most effective embedding and classification model/layers for these two tasks, although we also optimise additional hyper-parameters as well. Our results are reported in Table 7.1 (Text models) and Table 7.2 (Image models). The first three columns describe the setup of each model, while the remaining columns report the effectiveness and training time for that model on the two tasks (Informativeness and Humanitarian Categorization). The highest performing uni-modal model under eadch metric is highlighted in bold. We discuss our observations below:

Text - Word2vec vs. BERT: The first factor of interest is the model that we use to embed the text. We compare two approaches here, the shallow word embedding approach Word2Vec and the deep neural language model BERT-Base. Examining the F1 scores for both tasks in Table 7.1, we observe that BERT-Base consistently outperforms Word2Vec, as expected. Indeed, the best performing models for both tasks are BERT-Base model with 86.4% F1 for Informativeness and 83.2% for Humanitarian Categorization.

Text - Classification Layer Types: The second primarily variable that can impact performance is the classification model / layers that we use to convert the text embedding into a classification. We compare three different possible configurations: 1) the simplest approach of adding

			Informativeness Task					Humanitarian Categorization Task				
Image Embedding	Classification Layer	Batch size	Acc	precision	recall	F1	Time	Acc	precision	recall	F1	Time
VGG16	Dense(1)	16	0.8330	0.8310	0.8330	0.8320	00:19:46	0.7680	0.7640	0.7680	0.7630	00:20:28
ResNet50	Dense(1)	16	0.8130	0.8188	0.8160	0.8174	00:16:32	0.7420	0.7370	0.7410	0.7390	00:16:17
ResNet152	Dense(1)	16	0.8239	0.8219	0.8208	0.8213	00:08:40	0.7626	0.7530	0.7620	0.7575	00:09:13
ResNet152	Dense(1)	32	0.8388	0.8366	0.8353	0.8359	00:14:11	**0.7843**	**0.7692**	0.7712	0.7702	00:14:55
ResNet152	Dense(1)	64	0.8437	0.8331	0.8299	0.8315	00:19:47	0.774	0.7652	0.7791	**0.7721**	00:20:16
ResNet152	Dense(1)	128	**0.8573**	**0.8443**	**0.8437**	**0.8440**	00:31:44	0.7830	0.7340	**0.7960**	0.7637	00:12:15

Table 7.2: Crisis Content Categorization of Uni-Modal: **Image** models.

a single dense layer (**DO|Dense(1)**), 2) adding a deeper network comprised of multiple dense layers (**DO/Dense(3)**), and the classifier configuration used by [170] (**Norm/Conv(3)|Dense**). As we can observe from Table 7.1, the classifier that includes the convolutional layers is the most effective in nearly all cases. This is most notable for the Word2Vec based models, where the (**Norm/Conv(3)|Dense**) classifier resulted in an around 24% gain in Informativeness F1 performance. However, the gains narrow dramatically when using BERT-base as the embedding model, with between 0-6% F1 gains observed for Informativeness and 0-2% gains for Humanitarian Categorization in comparison to using dense layers only. Hence, we conclude that it is better to include convolutional layers within the classifier, although we note that this comes at a significant cost in increased training time.

Text - Optimizer: Finally, when training the text classifier, a lesser factor that can influence the model performance is the optimiser that is used during training, with two common ones being Adam and AdamW. We also report a comparison of these two optimiser in Table 7.1. From our results, we can conclude that the AdamW optimiser appears to be superior here, indeed in some scenarios it has a surprisingly large impact. For instance, for Humanitarian Categorization, simply changing the optimiser resulted in an increase in performance from 80.7% to 83.2%.

Image - VGG16 vs. ResNet: Moving to the image modality, we have fewer variables to consider, as the classification layer is held constant for images. The most influential factor then is how we embed the images. In this case, we compare the VGG16, ResNet50 and ResNet152 pre-trained image embedding models in Table 7.2. As we can see from the Table, in terms of F1 performance across the two tasks, there is little difference in performance between the pre-trained image models. For instance, for the Informativeness task, the difference in performance between the VGG16 and the best ResNet model is only 0.2%. Although we note that to achieve equivalent performance with ResNet as VGG16, we needed to use a larger batch size of 128 than VGG's 16, so out-of-the-box VGG16 appears to be a safer option.

Text vs. Image Efficiency: Finally, one of the practical considerations when building machine learned models is how long they take to train. This is more significant when working in a multi-modal space, as we need to train the different modalities. For all models Table 7.1 and Table 7.2 reports the training times for the text and image models, respectively. Generally, we observe that the training times for both text and image models is between 10-20 minutes on average. However, a notable exception here is the introduction of the convolutional layers in the text classifier, that markedly increases training time to around 2 hours. We note that this should not

			Informativeness Task					Humanitarian Categorization Task				
Modalities	Embeddings	Classification Layer	Acc	precision	recall	F1	time	Acc	precision	recall	F1	time
Text	BERT-Base	DO/Dense(3)	0.8651	0.8638	0.8651	0.8620	00:19:42	0.8304	0.8345	0.8304	0.8318	01:55:01
	Word2Vec	Norm/Conv(3)|Dense	0.8080	0.8100	0.8100	0.8090	00:15:27	0.7040	0.7000	0.7000	0.6770	00:25:00
Images	ResNet152	Dense(1)	0.8573	0.8443	0.8437	0.8440	00:31:44	0.7830	0.7340	0.7960	0.7637	00:12:15
	VGG16	Dense(1)	0.8330	0.8310	0.8330	0.8320	00:19:46	0.7680	0.7640	0.7680	0.7630	00:20:28
Text+Images	Word2Vec+VGG16	Concat|Conv(3)|Dense	0.8440	0.8410	0.8400	0.8420	01:40:05	0.7840	0.7850	0.7800	0.7830	01:58:25
	BERT-Base+ResNet152	Concat|Conv(3)|Dense	0.8728	0.8776	0.8743	0.8759	02:49:03	0.8220	0.8321	0.8225	0.8273	02:59:21
	BERT-Base+ResNet152	Concat|Norm|DO|DO/Dense(2)	**0.8977**	**0.8997**	**0.8977**	**0.8984**	00:26:04	**0.8670**	**0.8690**	**0.8670**	**0.8677**	00:54:12

Table 7.3: Comparison of Uni-Modal and Multi-Modal models for Crisis Content Categorization.

be a significant issue for a production system, but may slow down model development time if performing significant hyperparameter tuning.

Uni-Modal vs. Multi-Modal Comparison In the previous section we built optimised deep neural models for both CrisisMMD Informativeness and Humanitarian Categorization tasks. The best text-only models achieved 86.4% and 83.18% F1 respectively, while the best image-only models were slightly lower, achieving 84.4% and 77.2% F1 respectively. We now answer our core question: is multi-modal data key for these tasks? If so, by constructing multi-modal models we should be able to significantly enhance performance over these best uni-modal models.

Table 7.3 reports a performance comparison between the best uni-modal models and three late interaction multi-modal models. As before, we distinguish models based on the embedding approach and classification layer(s) used. The Word2Vec+VGG16 into a Concat|Conv(3)|Dense classification layer is the approach by [170], which we use as a baseline. From Table 7.3, we make the following observations. First, we can compare each multi-modal model with the uni-modal models that comprise it. Starting with the approach by [170], the multi-modal model (row 5) achieved 84.2% and 78.3% F1 under each task, respectively. This is a marked increase over the two uni-modal models that comprise it, i.e. row 2 (text-only with 80.1% and 67.7% F1) and row 4 (image-only with 83.2% and 76.3% F1), confirming results from their paper. On the other hand, this combined model is still quite weak in terms of overall performance in comparison to the best uni-modal models we created earlier. Hence, we next compare how using more effective embeddings can increase performance. If we simply replace Word2Vec and VGG16 with BERT-Base and ResNet152 the resulting model (row 6) exhibits much higher overall performance: 87.6% and 82.7% F1 respectively, which is more effective than the best uni-modal models for the Informativeness task, although it is still slightly less effective than the best uni-modal model for the Humanitarian Categorization task. However, one possibility is that the classification layer being used is not expressive enough. As such, we ask if we replace the convolution-based classifier with a more expressive dense layer architecture (row 7). As we can see, the transition to a more expressive classification layer results in a further marked increase in performance, to almost 90% F1 for the Informativeness task and 86.7% F1 for the Humanitarian Categorization task, which outperform the best uni-modal models for these tasks.

Task	# Test	Text+Images			Text-Only Failed	Image-Only Failed
	Tweets	Embeddings	Classification Layer	Outcome	(BERT-Base, DO/Dense(3))	(ResNet152, Dense(1))
Informativeness	1534	BERT-Base+ResNet152	Concat|Norm|DO|DO/Dense(2)	Correct	75 (4.9%)	228 (14.9%)
				Failed	126 (8.2%)	78 (5.0%)
Humanitarian Categorization	955			Correct	177 (11.5%)	199 (13.0%)
				Failed	96 (6.3%)	81 (5.2%)

Table 7.4: Comparison of the number of tweets classified correctly and incorrectly by the uni-modal and multi-modal models.

Overall, we have shown that combining evidence from multiple modalities can bring marked gains in performance, however our results have highlighted the importance of using strong embeddings (particularly for text) and an expressive classification layer, if state-of-the-art performances are to be achieved.

7.2.5 Additional Observations and Discussion

Having answered our core question, we next perform additional analysis on the uni-modal and multi-modal models produced to evaluate the strengths and weaknesses of these models. Indeed, just because a model overall has a higher performance does not mean that it is more effective in all cases. We divide our analysis into two components: 1) a failure analysis of the best models to see where the uni-modal and multi-modal models differ in terms of classification error distribution; and 2) provision and analysis of illustrative examples.

Error Distribution Analysis We begin by contrasting the success and failures of the best multi-modal model against the uni-modal models that comprise it. In particular, Table 7.4 reports for both Informativeness and Humanitarian Categorization tasks the number of test tweets the multi-modal model succeeded or failed to classify correctly, and the associated text-only and image-only models failed *also* to classify correctly. In effect, the 'Failed' Outcome rows report the number of cases where both the multi-modal and uni-modal models failed, while the 'Correct' Outcome rows report the number of cases where the multi-modal model succeeded, but the uni-modal models did not. Higher numbers in 'Correct' rows indicate errors that the uni-modal made that were fixed by the multi-modal model. The counts in the 'Failed' rows indicate tweets which neither model classified correctly (indicating that more work is needed and the scope for improvement available).

As we can see from 'Correct' rows in Table 7.4, the multi-modal model is correcting a large number of errors that the uni-models made. For the uni-modal text models, 4.9% and 11.5% of tweets were corrected, while for the uni-modal image models, 14.9% and 13% of tweets were corrected. This demonstrates again the value that integrating text and image evidence together can bring to crisis content categorization. On the other hand, we see that even with the best multi-modal model, there is still significant scope for improvement, with between 5% to 8.2% of tweets remaining incorrectly classified, even with the addition of multi-modal evidence.

To examine where these errors occur in more detail, Figure 7.2 visualises the confusion matrices for the different models for the Humanitarian Categorization task. The axes on the confusion matrices denote the primary categories in the task, namely: "Affected individuals" is denoted as "A"; "N" represents the "not-humanitarian" category; "I" to represents "infrastructure and utility damage"; "O" denotes "Other relevant information"; and "Rescue, volunteering or donation effort" is denoted as "R". The values contained within each box and associated colours represent the proportion of tweets in each pairing. The left-right centre diagonal represents correct classifications.

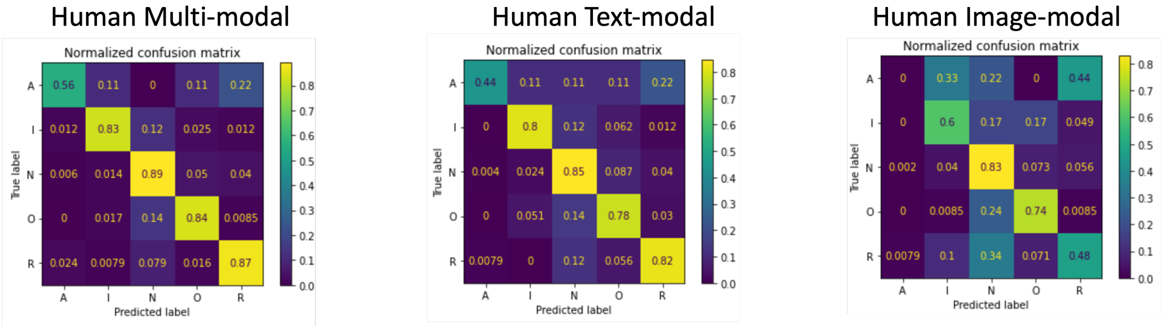
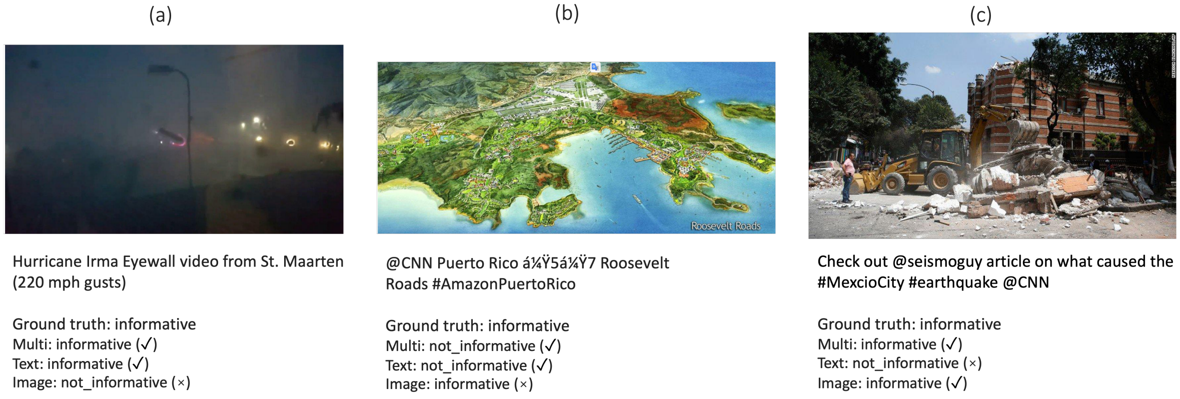


Figure 7.2: Confusion matrices for the uni-modal and multi-modal models. Due to the name of each class being too long, "Affected individuals" is denoted as "A"; "N" represents the "not-humanitarian" category; "I" to represents "infrastructure and utility damage"; "O" denotes "Other relevant information"; and "Rescue, volunteering or donation effort" is denoted as "R".

From Figure 7.2 we see that the multi-modal model improves performance across all 5 primary categories, however the distribution of these gains are not the same comparing across the uni-modal models. Comparing against the text-only model, consistent 3-6% improvements are observed across the categories, with the exception of the affected individuals category that has a larger 12% uplift. However, in contrast, the image-only model performs particularly poorly on two categories: affected individuals (0% correct) and Rescue, volunteering or donation effort (48% correct). As a result, the associated up-lift from the multi-modal combination is larger. It is also worth highlighting that the gains we are seeing here are not simply additive, i.e. the multi-modal model is able to classify tweets correctly that neither of the uni-modal models could. We can see this most clearly in the case of the affected individuals class, where the text model classified 44% of the tweets correctly and the image model classified 0% correctly, but the combination in the multi-modal model managed to classify a larger 56% of the tweets in this class correctly.

Analysis of Illustrative Classification Examples Finally, we provide several illustrative examples to provide insights into the success and failure modes of the multi-modal model. To begin, Figure 7.3 shows three examples where the multi-modal model correctly categorises tweets, but the uni-modal models do not. Examples (a) and (b) show cases where the image unimodal model failed, but the multi-modal model makes the right prediction. Specifically, they illustrate scenarios when the images' contents are vague or not informative, but the text message is very clear. Example (c) in contrast, is a tweet that the text model gave the wrong prediction, but the multi-modal model correctly classifies it through the use of features from the image.



(a) Hurricane Irma Eyewall video from St. Maarten (220 mph gusts)

Ground truth: informative
Multi: informative (✓)
Text: informative (✓)
Image: not_informative (×)

(b) @CNN Puerto Rico á¼Ÿ5á¼Ÿ7 Roosevelt Roads #AmazonPuertoRico

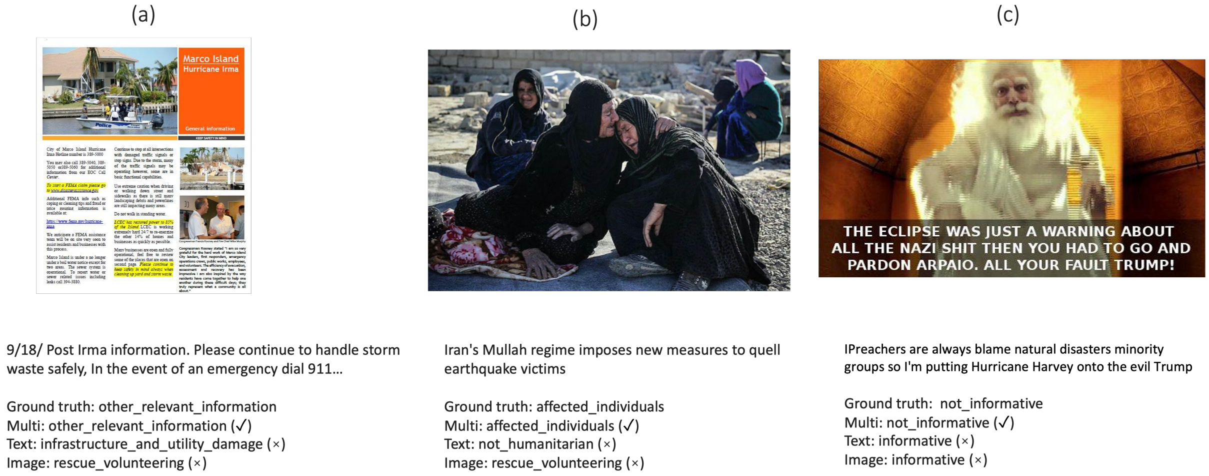
Ground truth: informative
Multi: not_informative (✓)
Text: not_informative (✓)
Image: informative (×)

(c) Check out @seismoguy article on what caused the #MexcioCity #earthquake @CNN

Ground truth: informative
Multi: informative (✓)
Text: not_informative (×)
Image: informative (✓)

Figure 7.3: Examples of multi-modal model filter misleading info in single modality.

Furthermore, as we noted in the error distribution analysis, the gains from the multi-modal model are not simply additive, i.e. it is learning how to relate both the text and image evidence together to gain a deeper understanding of the content than is possible from one modality alone. Figure 7.4 provides examples where this occurs, by showing tweets where both uni-modal models incorrectly classified the tweet, but the multi-modal model classified it correctly. For instance, in example (a), the text and image present a large volume of information, making it difficult for the model to determine the core subject of the tweet. Meanwhile, example (b) illustrates a case where the text and image present different information; the text indicates that there is no need for immediate rescue, but the image shows an injured person. Example (c) is a difficult tweet to classify, as the text contains many keywords that are related to crises, but the actual content is not relevant to the current crisis being investigated.


(a)
(b)
(c)
Marco Island
Hurricane Irma
General information
THE ECLIPSE WAS JUST A WARNING ABOUT ALL THE NAZI SHIT THEN YOU HAD TO GO AND PARDON ARPAIO. ALL YOUR FAULT TRUMP!

(a) 9/18/ Post Irma information. Please continue to handle storm waste safely, In the event of an emergency dial 911...

Ground truth: other_relevant_information
Multi: other_relevant_information (✓)
Text: infrastructure_and_utility_damage (×)
Image: rescue_volunteering (×)

(b) Iran's Mullah regime imposes new measures to quell earthquake victims

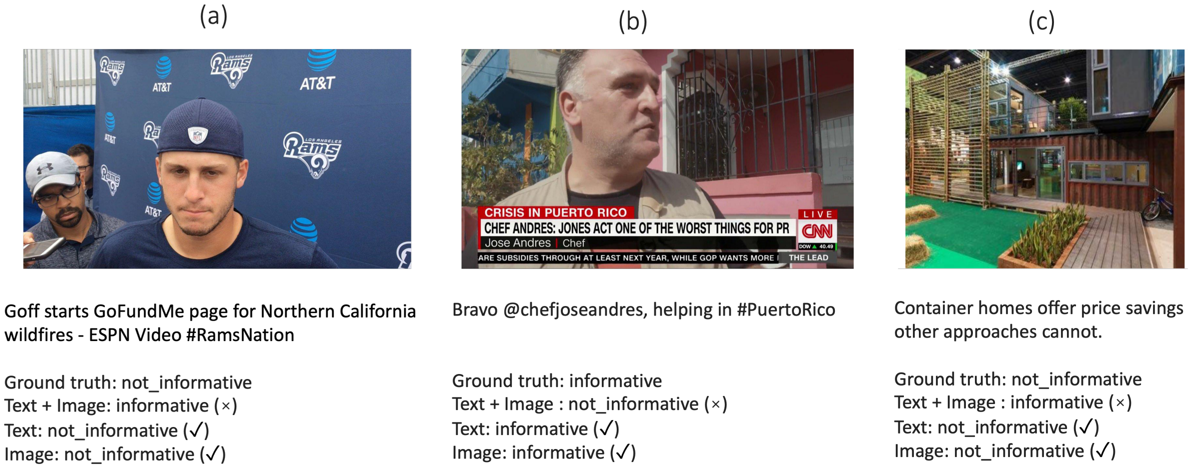
Ground truth: affected_individuals
Multi: affected_individuals (✓)
Text: not_humanitarian (×)
Image: rescue_volunteering (×)

(c) IPreachers are always blame natural disasters minority groups so I'm putting Hurricane Harvey onto the evil Trump

Ground truth: not_informative
Multi: not_informative (✓)
Text: informative (×)
Image: informative (×)

Figure 7.4: Examples of multi-modal model beats uni-modal model.

On the other hand, the multi-modal model is not always better than the uni-modal models, although these cases are rare (85 examples). From analysing these, there seem to be two broad failure modes. First failure mode is cases where the model appears to have been misled by examples where the true label is debatable. For instance, Figure 7.5, (a) and (b) describe a similar story of how a hero acted during the crisis, but assessors give different labels to them, making it likely that the model will learn the wrong patterns. Meanwhile, there are also a small number of labelling errors, such as example (c), where the assessor selected "not informative", but we believe this tweet should be "informative". The second broad scenario is cases where both the text uni-modal model and image uni-modal model give consistent predictions but the multi-modal model produces a different (incorrect) prediction, such as the examples shown in Figure 7.6 for the Informativeness task. We suspect that if the multi-modal model identifies a weak connection between the text and image, the model would judge that neither information within the text nor image to be informative, but this would require more analysis to determine definitively.



(a) Goff starts GoFundMe page for Northern California wildfires - ESPN Video #RamsNation

Ground truth: not_informative
Text + Image: informative (×)
Text: not_informative (✓)
Image: not_informative (✓)

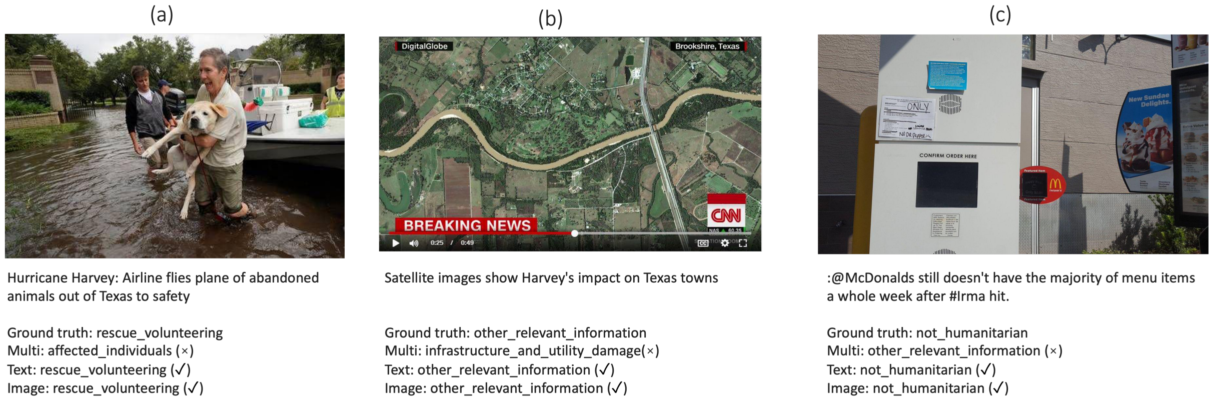
(b) Bravo @chefjoseandres, helping in #PuertoRico

Ground truth: informative
Text + Image : not_informative (×)
Text: informative (✓)
Image: informative (✓)

(c) Container homes offer price savings other approaches cannot.

Ground truth: not_informative
Text + Image : informative (×)
Text: not_informative (✓)
Image: not_informative (✓)

Figure 7.5: Examples of problem of true label in CrisisMMD dataset.



(a) Hurricane Harvey: Airline flies plane of abandoned animals out of Texas to safety

Ground truth: rescue_volunteering
Multi: affected_individuals (×)
Text: rescue_volunteering (✓)
Image: rescue_volunteering (✓)

(b) Satellite images show Harvey's impact on Texas towns

Ground truth: other_relevant_information
Multi: infrastructure_and_utility_damage(×)
Text: other_relevant_information (✓)
Image: other_relevant_information (✓)

(c) :@McDonalds still doesn't have the majority of menu items a whole week after #Irma hit.

Ground truth: not_humanitarian
Multi: other_relevant_information (×)
Text: not_humanitarian (✓)
Image: not_humanitarian (✓)

Figure 7.6: Examples of multi-modal model perform worse than uni-modal model.

7.2.6 Summary

Social media is a critical platform for emergency response agencies to acquire actionable information, but extracting information from these unstructured and multimodal sources poses significant challenges. As a step towards building more effective crisis response tools, we investigated the importance of considering all modalities within social media data. Specifically, we compared state-of-the-art unimodal and multimodal models to assess the benefits that multimodal models can bring.

Through experimentation on the CrisisMMD Informativeness and Humanitarian Categorization tasks, we demonstrated significant performance gains by fusing text and image evidence for crisis content categorization. We observed around a 6% gain in F1 performance for the Informativeness task and a 4% increase in Humanitarian Categorization F1 performance when moving from unimodal to multimodal models, as detailed in Section 7.2.4.

Moreover, our analysis in Section 7.2.5 showed that the multimodal model does not simply choose a unimodal model to apply on a case-by-case basis. Instead, it effectively fuses evidence from both modalities, allowing it to correctly classify examples that could not be accurately categorized by any unimodal model alone. We also provided an analysis of the success and failure modes of the multimodal model, offering insights into its areas of effectiveness.

These findings support our hypothesis and thesis statement by demonstrating that multimodal data is essential for accurately categorizing crisis content on social media. By integrating textual and visual information, our approach significantly improves the performance and reliability of crisis response tools, thereby validating the core claims of our research.

7.3 Conclusion

In this chapter, we addressed research question 3 proposed in Section 4.3 within the context of crisis response.

In Section 7.2, we undertake the automated classification of social media information through multimodal approaches, incorporating both vision and language modalities. We examined the importance of integrating multiple modalities for crisis content categorization, designing a pipeline for multimodal learning that fuses textual and visual inputs and classifies the content based on the specified task. Evaluating with the CrisisMMD dataset, we demonstrated effective automatic labeling for this task, achieving an average of 88.31% F1 performance across two significant tasks (relevance and humanitarian category classification). Most importantly, in Section 7.2.5, we analyzed the success and failure cases of unimodal and multimodal models. Our analysis showed that the MCA framework makes decisions based on information from both modalities, rather than relying on a single modality as previous multimodal approaches did. This indicates a deeper understanding of multimodal information, achieved through better embeddings and deeper intra-modal and inter-modal alignment, supporting our core hypothesis in the thesis statement.

In summary, we confirmed that utilizing multimodal data is the most effective approach for categorizing crisis content on social media. Our findings provide evidence of increased inter-modality alignment and enhanced performance metrics, such as accuracy and F1 score, thereby demonstrating the power and effectiveness of the proposed MCA framework. This supports our thesis statement in Section 1.2 by validating the hypothesis that improved alignment and integration of multimodal data lead to superior performance in crisis response tasks.

Chapter 8

The Applications of the Proposed MCA Framework in Robotic Vision

8.1 Introduction

In this chapter, we aim to evaluate whether our proposed multimodal learning framework, MCA, improves performance in the context of robotic vision by addressing the intra-modal alignment issues outlined in the thesis statement (see Section 1.2). We expect to see improvements in performance metrics such as accuracy and false positive rate, thereby demonstrating the effectiveness of our MCA framework.

Recent advances in deep learning have led to the emergence of Multimodal (vision + language) Large Language Models (MLLMs) such as BLIP2 [123] and BEiT-3 [252]. These MLLMs, trained on large-scale datasets, act as general-purpose encoders for multiple modalities and offer substantial potential for transfer to various applications. While they have shown state-of-the-art performance in fields such as conversational agents and information retrieval, their utility in robotic vision remains largely unexplored. Early studies, like PaLM-E [50], have begun to delve into this area but primarily focus on vision and language understanding, largely overlooking the unique real-world challenges of robotic vision, such as task-specific camera poses, variable lighting, and clutter. Therefore, prompted by the success of our proposed MCA framework in other domains, this chapter addresses the question: *To what extent can MLLMs improve vision tasks specific to the robotics domain?*

The recently introduced ARMBench dataset [161] from Amazon exemplifies the complexities inherent to robotic vision. It comprises three critical robotic vision perception tasks: object instance segmentation, object identification, and defect detection, and presents a large-scale representation of real-world scenarios—specifically, object grasping and manipulation tasks in Amazon warehouses. These tasks demand a vision system capable of handling an extremely large range of objects (e.g., 190K+ unique objects), robust to variable lighting conditions, and capable of operating effectively in cluttered environments. Moreover, the ever-changing inventory of warehouses requires the vision system to perform well in transfer learning scenarios. The individual tasks, though extensively researched, have yielded task-specific models, complicating their integration into a unified vision pipeline. Such integration involves selecting the right model for each task, coordinating them effectively, and fine-tuning each one – a process that is both time-consuming and challenging in terms of engineering.

We argue that MLLMs have better transfer capability compared to previous task-specific models, owing to MLLMs' large-scale pre-training, which allows them to serve as robust backbones for these tasks. This substantially reduces the engineering complexity of developing different backbones for multiple tasks and the time required for model selection and tuning. Additionally, MLLMs have shown particular resilience to out-of-distribution examples, such as objects presented in novel poses or amidst visual distractions, which are critical issues in robotic vision perception tasks.

In this chapter, we introduce RoboLLM, based on our proposed multimodal learning framework, MCA from Chapter 4. This framework is designed to adapt MLLMs to the multifaceted challenges of robotic vision, using the ARMBench dataset for evaluation. Our main contributions are as follows:

- We present RoboLLM, a generalized framework that employs pre-trained MLLMs as backbones for tackling complex robotic vision tasks, including object instance segmentation, object identification, and defect detection. This supports our thesis statement by demonstrating the versatility and effectiveness of our proposed multimodal learning framework.

- We are the first to address all three key vision tasks in the ARMBench dataset, which represents challenging, large-scale robotic manipulation scenarios. This comprehensive approach supports our hypothesis that a robust multimodal framework can handle diverse and complex tasks.

- We introduce a lightweight variant of the BEiT-3 architecture aimed at increased performance and efficiency (Section 8.3.2), broadening its applicability to resource-constrained robotic applications. This aligns with our thesis objective of improving the efficiency of multimodal learning frameworks.

- Our experiments (Section 8.3.6) show that RoboLLM achieves state-of-the-art results across all three ARMBench tasks, validating our hypothesis that integrating our multimodal learning framework, MCA, enhances performance in robotic vision tasks.

- We further demonstrate RoboLLM's robustness to object number variance and its superior performance on out-of-distribution examples in the object segmentation task, where previous works fail. Notably, RoboLLM achieves a 97.8% recall@1 in the object identification task. These results support our thesis statement by proving that improved intra-modal and inter-modal alignment leads to better performance and generalization.

- The original material in this section has been accepted for presentation at the 2024 IEEE International Conference on Robotics and Automation, a Core ranking A* conference with an h5-index of 119.

8.2 Revisiting Transformers vs CNNs

Large language models (LLMs) predominantly utilize transformers rather than convolutional neural networks (CNNs) due to architectural differences that make transformers better suited for natural language processing (NLP) tasks. The sequential nature of language, global context requirements, scalability needs, and the specific mechanisms of transformers collectively establish them as the architecture of choice for LLMs.

Language inherently involves sequential data where the meaning of a word depends on its context within a sentence or passage. Transformers excel in processing such sequences through the self-attention mechanism, which enables each token in a sequence to attend to all others simultaneously. This global context awareness ensures that transformers capture both local and distant relationships efficiently, a feature critical for understanding complex language structures. In contrast, CNNs, designed for grid-like data such as images, rely on localized receptive fields and lack the innate ability to handle long-range dependencies effectively. While CNNs can process sequential data through 1D convolutions, their reliance on fixed or expanded receptive fields limits their capacity to model global relationships in long sequences.

Transformers also address the challenge of long-sequence modeling through positional encodings and multi-head attention. Positional encodings explicitly incorporate sequence order into the model, allowing the self-attention mechanism to flexibly model relationships between distant tokens. Furthermore, transformers process sequences in parallel, significantly enhancing training efficiency on GPUs and TPUs. This parallelism contrasts with the inherently sequential nature of CNNs when applied to ordered data, which leads to slower training times, particularly for tasks involving extensive sequences.

The empirical success of transformers further cements their role in NLP. Models like GPT, BERT, and T5, built on transformer architecture, consistently achieve state-of-the-art performance across a wide range of tasks, from text generation to question answering. These models leverage the self-attention mechanism to capture deep contextual relationships, outperforming previous approaches. Attempts to adapt CNNs for NLP tasks, such as TextCNN, have demonstrated potential in small-scale scenarios but fail to scale effectively or capture the nuanced dependencies required for large-scale language modeling.

Transformers also exhibit remarkable flexibility for multimodal applications, such as combining vision and text. Frameworks like CLIP and Vision Transformers demonstrate how trans-

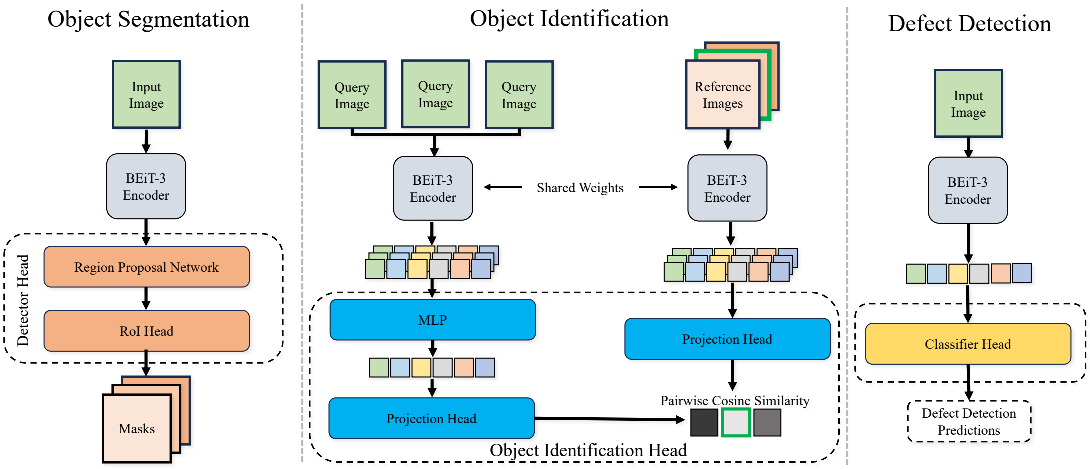


Figure 8.1: The task specific adaptions to BEiT-3, proposed in our RoboLLM framework, for tackling each challenge in ARMBench. BEiT-3's large-scale vision-language pretraining allows it to be easily and effectively transferred to downstream tasks.

formers can seamlessly integrate data from multiple modalities. In contrast, CNNs, while excelling at processing spatial data like images, often require complex hybrid architectures or preprocessing pipelines to accommodate text and other modalities.

In conclusion, the dominance of transformers in LLM design stems from their ability to capture both local and global context through self-attention, handle long sequences efficiently, and support parallelized training. Their adaptability for multimodal data and proven success in state-of-the-art models further reinforce their suitability for large-scale NLP applications over CNNs.

8.3 RoboLLM: Robotic Vision Tasks Grounded on Multimodal Large Language Models

8.3.1 Overview

The proposed framework, RoboLLM, is designed to address three primary vision tasks in robotic manipulations as defined in the ARMBench dataset: *object segmentation, object identification, and defect detection*. As illustrated in Figure 8.1, the architecture of RoboLLM is inherently modular. This modular design allows for the integration of a variety of Multimodal Large Language Models (MLLMs) as backbone encoders, alongside task-specific heads tailored for distinct robotic vision tasks. This design principle of decoupling the backbone from specific downstream tasks offers several advantages, including ease of maintenance and flexibility for quick and easy adaptation to new vision tasks, while fully exploiting the benefits of large-scale pre-trained models. Task-specific heads are integrated to tackle each unique vision challenge, with the choice of heads justified by their proven effectiveness in the respective tasks. Subse-

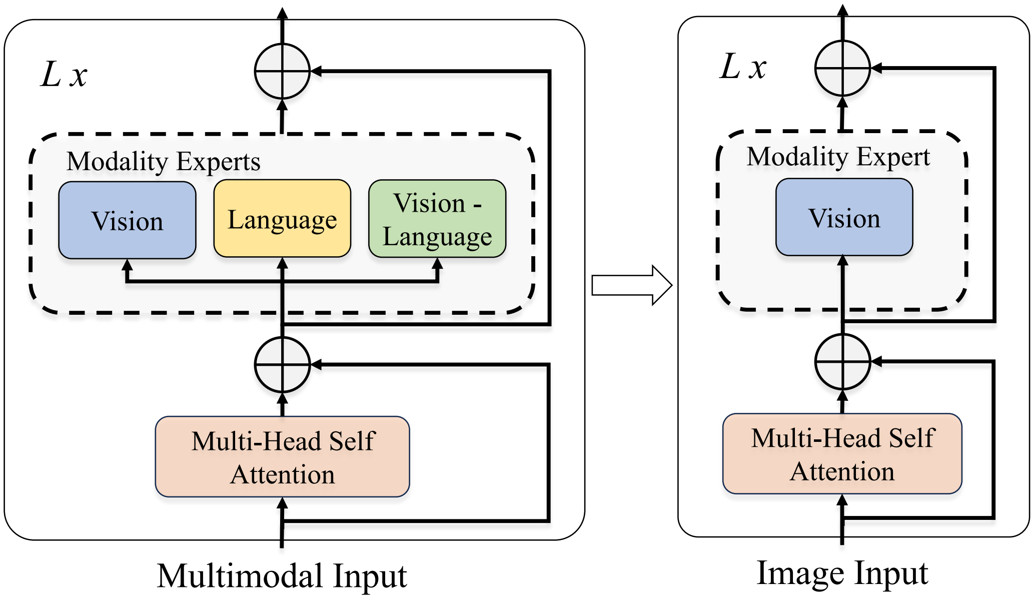


Figure 8.2: Our lightweight modification of the BEiT-3 [252] backbone only remains the vision experts.

quent sections elaborate on the role of the backbone encoder and provide specific configurations for each vision task in the ARMBench dataset.

8.3.2 Backbone Encoder

The backbone encoder is a critical component in the RoboLLM framework, which is responsible for producing feature maps for downstream tasks. It is based on a widely-used MLLM architecture featuring a MultiWay Transformer design [252]. Our design philosophy allows the backbone to be replaced with future MLLM developments, making the system future-proof in the rapidly evolving domain of robotic vision. Currently, we believe BEiT-3 is the most suitable option owing to its state-of-the-art performance in various vision benchmarks [252]. Moreover, BEiT-3 demonstrates that features learned in a multi-modal setting are more task-agnostic and generalized, thereby offering better transferability to different vision tasks without the need for extensive fine-tuning. These characteristics of BEiT-3 make it highly suitable for robotic vision tasks.

The MultiWay Transformer blocks, illustrated on the left side of Figure 8.2, comprise a shared self-attention module and a pool of feed-forward networks known as modality experts. These experts are designed for handling different types of data. However, this paper is focused only on addressing robotic vision perception tasks, thereby eliminating the need for other modality experts besides vision. Thus, our streamlined architecture, as shown on the right side of Figure 8.2, eliminates non-vision modality experts, reducing both computational overhead and the parameter count from 222 million to 87 million in the BEiT-3 Base model [252]. This

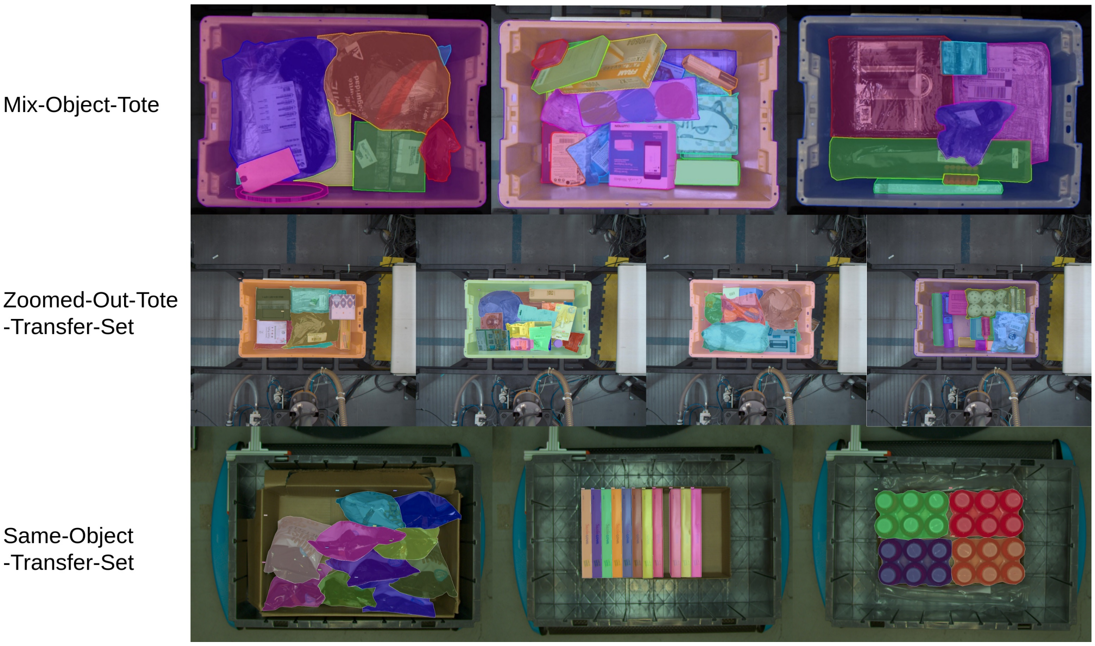


Figure 8.3: Examples for different image segmentation tasks in ARMBench [161].

design choice not only improves computational efficiency but also retains the flexibility to incorporate other modalities in future iterations of the task. Importantly, the feature maps generated by the backbone are exploited across all three vision tasks addressed in this work, demonstrating the framework's effectiveness and efficiency.

8.3.3 Object Segmentation

Object segmentation serves as the prerequisite task within the ARMBench dataset, aiming to identify and delineate individual objects in containers.

Semantic segmentation is a dense prediction task in computer vision that involves classifying every pixel in an input image. Recent state-of-the-art methods often utilize Fully Convolutional Networks (FCNs), which consist of a deep convolutional neural network serving as the encoder (or backbone) and a decoder tailored for segmentation to produce dense predictions. More recently, Vision Transformers (ViTs), which leverage a spatial attention mechanism, have been introduced to computer vision tasks. Unlike traditional convolution-based backbones, ViTs employ a straightforward, non-hierarchical architecture that maintains the resolution of feature maps throughout the network. This absence of down-sampling processes (aside from initial image tokenization) introduces distinct architectural considerations when using a ViT backbone for semantic segmentation.

Aligned with RoboLLM's overarching philosophy of decoupling the backbone from downstream tasks, we employ a plain-backbone approach for object segmentation based on ViT design. Unlike most object detection approaches, this plain-backbone detection approach does not require any pretraining on detection tasks, thereby eliminating hierarchical constraints on the backbone. This is a necessary step given the absence of a segmentation task during MLLM pretraining. Specifically, we transfer MLLM to object segmentation tasks, with modifications performed exclusively during the fine-tuning stage. The rationale behind this choice is to maximize the benefits accrued from pre-training on large-scale datasets, which in turn improves segmentation performance. Thus, to translate the capabilities of MLLM into practical application for object segmentation, a task-specific head is crucial for integrating and adapting the backbone's generic feature maps to the specificities of the task at hand.

Feature Pyramid Construction The cornerstone of our object detection strategy is the construction of a feature pyramid, following [130]. For this, we utilize only the last embeddings from the backbone, which is hypothesized to contain the most discriminative features. We then execute a series of parallel convolutions and deconvolutions to generate multi-scale feature maps. Specifically, starting with the default Vision Transformer (ViT) embeddings with a scale of $\frac{1}{16}$, we produce feature maps at scales of $\frac{1}{32}, \frac{1}{16}, \frac{1}{8}, \frac{1}{4}$ via convolutional strides of $2, 1, \frac{1}{2}, \frac{1}{4}$, respectively.

Task-Specific Head The task-specific head for object segmentation integrates a detector based on a Cascade Mask R-CNN [20]. This modular design facilitates seamless interaction with the feature maps generated by the backbone and is extensible for compatibility with other detector heads.

These generated multi-scale feature maps are processed through the Cascade Mask R-CNN detector head. A Region Proposal Network (RPN) from Cascade Mask R-CNN proposes candidate object bounding boxes, followed by mask generation for each object via a Region-of-Interest (ROI) head aslo from Cascade Mask R-CNN, which extracts pertinent features from the RPN.

Advantages and Rationale Our design offers multiple benefits. Firstly, the plain-backbone approach necessitates only modest modifications during the fine-tuning stage, which preserves the generality and extensibility of the pre-trained MLLMs. Secondly, the task-agnostic nature of our backbone allows for seamless interchangeability with various detector heads, providing flexibility in addressing a broad range of object segmentation tasks. Lastly, the selective use of only the last feature map for object detection aims to utilize the most potent features, thereby enhancing the system's overall efficacy and efficiency.

8.3.4 Object Identification

Object identification is the second task in the ARMBench dataset. This task is concerned with precisely categorizing detected objects among a database of predefined classes. In contrast to conventional identification and classification methods, which face computational difficulties

```
import numpy as np
# Q[n, h, w, c] - minibatch of query images
# R[n, h, w, c] - minibatch of reference images
# W_i[d_i, d_e] - learned proj of image to embedding
# labels - labels of whether Query image and
# reference image are the same class

# extract feature representations of each modality
if using both pre-pick and post-pick images:
    # BEiT-3 Encoder
    Q_f = encoder(Q) #[n, 3 x d_i]
    # Feature Aggregation
    Q_f = MLP(Q_f) #[n, 3 x d_i] to [n, d_i]
else:
    Q_f = encoder(Q) #[n, d_i]

R_f = encoder(R) #[n,d_i]
# project embedding [n, d_e]
Q_e = l2_normalize(np.dot(Q_f, W_i), axis=1)
R_e = l2_normalize(np.dot(R_f, W_i), axis=1)

# scaled pairwise cosine similarities [n, n]
# t - learned temperature parameter
logits = np.dot(Q_e, R_e.T) * np.exp(t)
# symmetric loss function
loss_Q = cross_entropy(logits, labels, axis=0)
loss_R = cross_entropy(logits, labels, axis=1)
loss = (loss_Q + loss_R)/2
```

Figure 8.4: Numpy-like pseudo-code for the core of an implementation of our framework for Object Identification task.

with numerous categories (exceeding 190,000 in the Amazon warehouse context), due to the need for an extremely large dense layer. Additionally, even if you manage to construct such a network, it struggles to accommodate the daily fluctuations in item numbers caused by Amazon warehouse operations, such as the introduction of new products or the depletion of stock. Therefore, we tackle this task as an image retrieval problem.

Task Variants In robotic manipulation within the context of the ARMBench dataset, the object identification task holds significance in pre- and post-object manipulation. In the pre-pick stage, object identification permits the retrieval of historically acquired objects or attributes for manipulation planning. While in the post-pick, the object's unique identifier is crucial for quality control and subsequent tasks. Thus, both pre- and post-pick images of the object could serve as query images, while the reference images in the container manifest act as gallery images. The challenge lies in accurately matching the query images to the gallery images.

Therefore, this task offers two variants: one reliant on pre-pick images (one for each pick) and the other incorporating post-pick images (two more for each pick). While the latter images present a greater challenge due to differing perspectives, object poses, and presentations, they enable the incorporation of multi-view data, thereby enhancing the overall retrieval performance.

Overall Architecture Consistent with our object instance segmentation task strategy, we employ a backbone plus a task-specific head architecture as shown in Fig. 8.1-center. For the two aforementioned task variants, the task-specific head comprises a projection head or adds a Multi-Layer Perceptron (MLP) for fusing pre- and post-pick images. A contrastive learning objective is applied to fine-tune the RoboLLM to optimize detection performance.

Contrastive Learning Fine-tuning We argue that high-quality feature maps suffice for measuring the similarity between query and reference images using dot product calculations. We maintain computational efficiency by abandoning complex retrieval techniques such as re-ranking, which is crucial for real-time robotic applications. Thus, a contrastive loss, InfoNCE [172], is used in the fine-tuning stage to improve the quality of the generated feature maps. It aims to minimize the distance between the feature maps of positive pairs (the same class) and maximize the distance between negative pairs (different classes), thus enhancing the efficacy of our image retrieval approach.

The backbone encoder transforms an input image or three images input Q into an output representation Q_f, which is further processed by a linear projection head to produce the final feature vector. In cases involving post-pick images as query images, their feature maps are concatenated prior to input into the MLP, providing a fused representation before feeding into the linear projection head. An $L2$ normalization is applied to Q_f to mitigate the risk of numerical instability during training. Specifically, given a batch of N (query image, reference image) pairs, our framework is trained to predict which of the $N \times N$ possible (query image, reference image) pairings across a batch actually occurred. To do this, our framework maximizes the pairwise cosine similarity of the query and reference image feature maps of the N real pairs in the batch, while minimizing the cosine similarity of the feature maps of the $N^2 - N$ incorrect pairings. We optimize a symmetric cross-entropy loss over these similarity scores. In Figure 8.4, we include the pseudocode of the core of the framework for the object identification task.

Advantages and Rationale This approach presents multiple advantages. First, the scalability of the image retrieval method sidesteps the challenges tied to implementing a large-category linear classification layer. Second, the modular design allows task-specific heads to be easily interchanged, making the framework adaptable to different identification scenarios. Third, computational efficiency is assured through the use of a dot product similarity measure and the elimination of intricate retrieval techniques. Lastly, the system's flexibility is highlighted by its ability to adapt to a variety of object identification scenarios, whether involving single or multiple query images.

Task	Mixed Object Tote		Zoomed Out Tote		Same Object Tote	
Model	mAP50	mAP75	mAP50	mAP75	mAP50	mAP75
ResNet50 + Mask RCNN *	0.72	0.61	0.25	0.19	0.11	0.10
RoboLLM	**0.82**	**0.67**	**0.57**	**0.45**	**0.15**	**0.13**

Table 8.1: Mean Average Precision at IoU thresholds of 50 and 75 across the different segmentation task subsets. * indicates results obtained from [161].

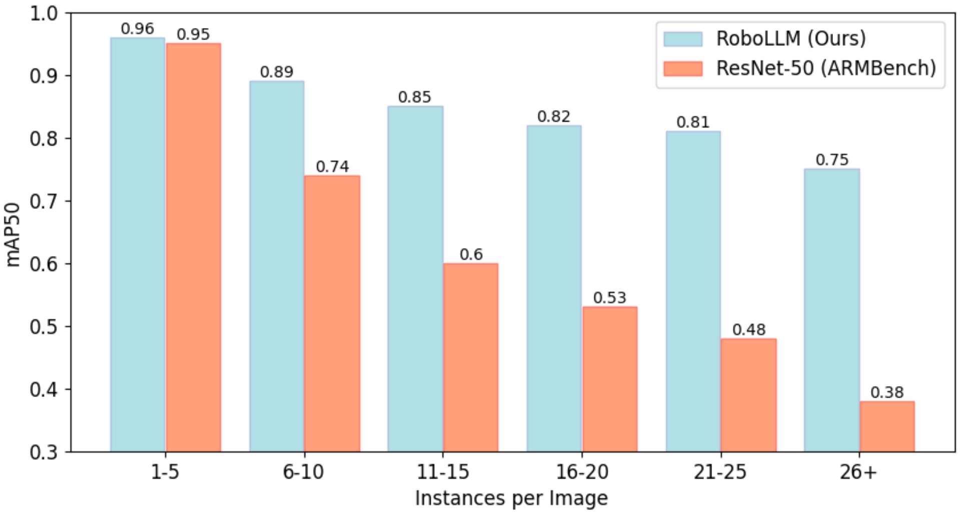


Figure 8.5: Number of object instances per image against Mean Average Precision at 50 on the Mixed-Object tote test-set.

8.3.5 Defect Detection

The third task in the ARMBench dataset is defect detection, aimed at identifying defects resulting from specific robotic manipulation activities. This is a crucial task as it directly impacts the integrity of the workflow and the quality of the end product. The dataset includes two types of robot-induced defects: 1) multi-pick, where multiple objects are mistakenly picked and transferred from the source to the destination container; and 2) package-defect, indicating activities that result in the object's packaging opening or the object deconstructing into multiple parts.

Consistent with the design strategy for earlier tasks, we employ a backbone plus a task-specific-head architecture. We reuse the same backbone encoder employed in the previous tasks. Utilizing the same backbone across multiple tasks ensures a cohesive and streamlined architecture. As in the object identification task, we leverage segmentations to locate objects. Unlike the object identification task, which has many categories, this task comprises only three classes: two types of defects and a nominal type. Therefore, a classification head is sufficient to conduct the classification and is appended to the backbone encoder for the purpose of making predictions across these three categories. We opt for a standard cross-entropy loss function for training, which is particularly suitable for categorical classification tasks.

Model	Ref Set	Recall@1		Recall@2		Recall@3	
		N=1	N=3	N=1	N=3	N=1	N=3
ResNet50-RMAC	Container	71.7	72.2	81.9	82.9	87.2	88.2
DINO-ViT-S	Container	77.2	79.5	87.3	89.4	91.6	93.5
BEiT-3-Base*	Container	83.7	84.5	83.8	N/A	84.5	N/A
RoboLLM	Container	**97.8**	**98.0**	**97.9**	**98.1**	**98.0**	**98.2**
RoboLLM	All refs	74.6	78.2	82.6	85.7	85.3	89.10

Table 8.2: Results on object identification at varying recall@k. * indicates no ARMBench fine-tuning. N=1 uses one pre-pick image, while N=3 uses three images per pick. “Ref set” specifies if reference images are container-specific or from the entire dataset.

8.3.6 Experiments

Experimental Setup Our experiments target the ARMBench dataset, focusing on three key robotic vision perception tasks: object instance segmentation, object identification, and defect detection. The experiments are conducted using two state-of-the-art Multi-Modal Learning Models (MLLMs) as the backbones of the proposed RoboLLM framework, namely BEiT-3 Base and BEiT-3 Large. Notably, the BEiT-3 Large model is employed exclusively for the defect detection task because the performance of the BEiT-3 Base model fails to meet the ideal task-specific requirements. We utilize the test sets of all tasks to report metrics. The backbones are initialized with pre-trained weights. A batch size of 64 is employed for all three tasks, with fine-tuning conducted over 50 epochs. An initial learning rate of 2×10^{-4} is set for all tasks. Early stopping is implemented if there is no improvement in the metrics on the validation set over a span of 5 epochs. For the object identification task, the Multi-Layer Perceptron (MLP) consists of two linear layers, with middle dimensions set to 3072 for the Base model and 4096 for the Large model. The projection head is implemented as a single linear layer, with its embedding dimension matching that of the encoder.

Object Segmentation Object segmentation is a critical stage within the robotic vision pipeline, influencing robotic planning, grasp operations, and object identification. Poor object instance segmentation can introduce defects during robotic manipulation, which is highly undesirable. The ARMBench challenge offers three data subsets to evaluate image segmentation performance: 1) *Mixed-Object-tote* serves as a general benchmark. 2) *Zoomed-out-tote* assesses model generalization to new warehouse environments. 3) *Same-Object* examines segmentation performance on tightly packed instances of identical objects.

We evaluate performance using mean-average-precision at Intersection-over-Union (IoU) thresholds of 0.5 (mAP50) and 0.75 (mAP75) across all subsets, as presented in Table 8.1. Our RoboLLM framework significantly outperforms a ResNet-50 baseline across all tasks: 1) On the *Mixed-Object-tote*, RoboLLM achieves 82% and 67% for mAP50 and mAP75, marking a 10%

and 6% improvement over ResNet-50. 2) The performance for RoboLLM on the *Zoomed-out-tote* is much better than it of ResNet-50, demonstrating superior generalization capabilities. 3) *Same-Object* poses the most difficult segmentation challenge, yet BEiT-3 still improves mAP50 by 4% over ResNet-50.

Furthermore, high variation in object instances between containers is common. To investigate model robustness under varying numbers of objects within an image, we report mAP50 in *Mixed-Object-tote* for different numbers of object instances (Figure 8.5). While both RoboLLM and ResNet-50 perform similarly at fewer than five instances (approximately 95% mAP), performance for ResNet-50 degrades significantly (to 38%) when the number of instances exceeds 26. In contrast, RoboLLM demonstrates a modest decline, maintaining 75% mAP for instances exceeding 26. This evidences that RoboLLM is robust for segmenting a large number of objects within an image.

Object Identification The object identification necessitates categorizing segmented objects to facilitate robotic planning for subsequent maneuvers. We use Recall@k as the evaluation metric, and the resultant performances are reported in Table 8.2. Using the pre-trained BEiT-3-Base model as a strong baseline, we observe an uplift in recall@1 performance from 77.2% to 83.7%, outperforming the best previously reported result using DINO-ViT-S [161]. Employing our RoboLLM framework, recall@1 is further improved to 97.8% and 98.0% when using only pre-pick images and both pre/post-pick images, respectively. This is a 21% increase over the best prior result [161], effectively solving ARMBench's object identification task.

Performance Under More Challenging Conditions We posit that this superlative performance could partially be credited to the specific challenge design, which restricts reference images to objects known to be present in the container. To evaluate RoboLLM in a more generalized scenario, we expand the set of reference images to include all unique objects within the dataset (190k+). As shown on the bottom line of Table 8.2, even under this more demanding setting, RoboLLM maintains an impressive 89.1% at recall@3. This robust performance suggests that our framework can adeptly handle even more complex, large-scale retrieval problems than those posed by the ARMBench challenge. We also find that our framework benefits from including additional query images in this challenging setting. This corroborates our design choice to aggregate multiple query images into a single representation for retrieval, thereby enhancing the model's robustness and versatility.

Defect Detection The practical deployment of automated defect detection in commercial warehouse settings presents substantial challenges, given the high demand of the performance. To contextualize, the ARMBench challenge outlines ideal performance criteria, requiring a recall rate exceeding 0.95 and an FPR below 0.01. Our results for precision, recall, and FPR are reported in Table 8.3. Our experiments demonstrate significant performance enhancements over

Task	Multi-Pick			Package Defect			Combined		
Model	Precision	Recall	FPR	Precision	Recall	FPR	Precision	Recall	FPR
ResNet50 *	-	0.34	0.05	-	0.73	0.05	-	0.57	0.05
RoboMLLM-B	**0.84**	0.98	**0.04**	**0.94**	0.91	0.04	**0.90**	0.94	**0.03**
RoboMLLM-L	0.82	**1.00**	**0.04**	0.89	**0.95**	**0.03**	0.86	**0.97**	**0.03**

Table 8.3: Metrics for defect detection tasks with RoboMLLM B and L backbones. * denotes [161] results. Combined metrics in right column. ARMBench ideal performance: $recall > 0.95$, $FPR < 0.01$.

a ResNet-50 baseline. Specifically, our RoboLLM with BEiT-Base achieves a combined recall rate of 94% across both types of defect detection, marking a 37% improvement over the baseline. Additionally, the combined FPR is reduced from 0.05 to 0.03. Despite these significant improvements, the system does not entirely meet the high recall and FPR criteria set forth by the ARMBench challenge.

To meet the desired performance set by ARMBench challenges, we conduct further experiments with a more powerful and larger model, BEiT-3 Large. Benefiting from our versatile yet robust framework, it is easy to adopt more powerful backbones as needed. Our results show that this model attains a combined recall rate of 97%, surpassing the desired recall target. However, the FPR remains at 0.03, indicating room for future enhancements in effectiveness. Overall, our RoboLLM framework shows significant improvements over existing methods, while a gap in the FPR warrants further investigation in future work.

8.4 Limitations of RoboLLM

Although RoboLLM demonstrates significant advancements in multimodal large language models for robotic tasks, it has notable limitations that warrant discussion. One major challenge lies in its domain-specific generalization capabilities. RoboLLM performs exceptionally well in tasks for which it has been trained or fine-tuned, such as ARMBench benchmarks. However, its ability to generalize to entirely new and diverse robotic domains or unforeseen scenarios remains constrained. This limitation stems from the lack of comprehensive coverage of niche or emerging robotic applications in its pre-training data, coupled with the model's reliance on its training distribution, which often results in performance degradation when faced with novel environments.

Real-time performance also presents significant obstacles for RoboLLM, particularly in the context of real-world robotic systems where latency and computational efficiency are critical. Due to its large-scale architecture, RoboLLM suffers from higher inference times, which can impede real-time decision-making. Additionally, its dependence on high-performance hardware, such as GPUs or TPUs, poses challenges for deployment on resource-constrained robotic platforms. Experimental results show that RoboLLM incurs an inference latency of approximately 7.3 milliseconds per image when tested on an Nvidia RTX 3060, underscoring its reliance on advanced computational infrastructure.

Another limitation arises from the cost associated with RoboLLM's development and deployment. The fine-tuning of such a large model requires substantial computational resources, which may be prohibitive for smaller organizations or academic institutions. Furthermore, the maintenance of RoboLLM, including updates and adaptations to accommodate new robotic systems or tasks, contributes to ongoing operational expenses.

The model's heavy dependence on the quality and diversity of its pre-training data also affects its robustness. Biases or gaps in the training dataset can result in blind spots in the model's reasoning, making it less effective in domains that are underrepresented in its training data. This dependence limits RoboLLM's adaptability, particularly for tasks outside its predefined training scope.

Addressing these limitations requires targeted research efforts. Optimizing model architectures could significantly reduce latency and computational demands, making RoboLLM more suitable for real-time applications. Additionally, improving interpretability and debugging tools for large language models in robotics would enhance their usability and reliability. By addressing these challenges, RoboLLM could achieve greater applicability and become a more versatile tool across a wider range of robotic domains.

8.5 Conclusions

In this chapter, we addressed research question 3 proposed in Section 4.3 within the context of robotic vision. We introduced RoboLLM, based on our proposed MCA framework, a highly efficient and effective framework designed to improve performance by enhancing intra-modal alignment, as claimed in our thesis statement (see Section 1.2). RoboLLM aims to establish a unified robotic vision pipeline that reduces engineering efforts, utilizing our proposed MCA framework as backbone encoders. We evaluated our framework on three distinct visual perception tasks: object segmentation, object identification, and defect detection, using the newly released Amazon ARMBench dataset [161], which is representative of large-scale real-world robotic vision problems.

Our results demonstrate that RoboLLM significantly outperforms previous benchmarks across all three challenges. Specifically, RoboLLM benefits from inherent knowledge gained from pre-training on large-scale multimodal data, requiring only a minimal task-specific head for each task. This approach not only significantly increases performance but also mitigates the engineering challenges associated with task-specific designs.

Notably, evidence of increased intra-modal alignment is observed through higher cosine similarity scores for images of the same class, indicating that contrastive learning has effectively enhanced embedding quality. Furthermore, we observed improved performance metrics, such as higher recall rates, which underscore the robustness and accuracy of the RoboLLM framework. All of this evidence supports our thesis statement in Section 1.2 by validating our hypothesis that improved alignment and integration of multimodal data lead to superior performance in robotic vision tasks.

The modular design of RoboLLM facilitates the incorporation of powerful backbones and task-specific modules, allowing for further performance enhancements as needed. Due to its versatility and effectiveness, RoboLLM serves as a highly practical and general solution to real-world robotic vision problems, demonstrating the applicability and success of our proposed multimodal learning framework, MCA, discussed in Chapter 4.

Chapter 9

The Applications of Proposed Multimodal Framework in Cross-modal Retrieval

9.1 Introduction

In this chapter, we aim to optimize our proposed multimodal learning framework, MCA, for cross-modal retrieval tasks by addressing the intra-modal and inter-modal alignment issues outlined in the thesis statement (see Section 1.2). We anticipate improvements in performance metrics such as accuracy and F1 score, thereby demonstrating the effectiveness of our MCA framework.

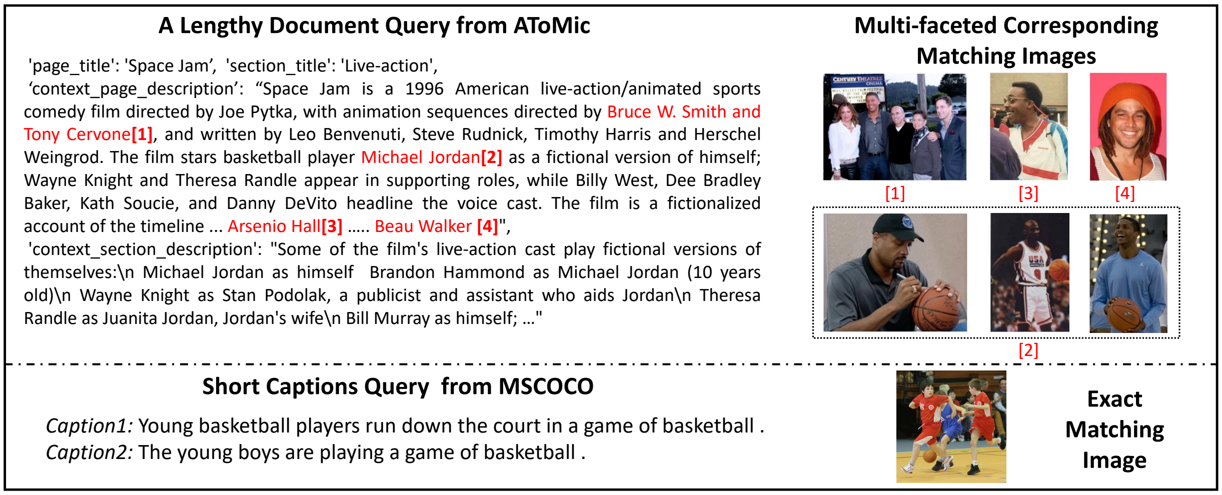
Text-to-image retrieval aims to locate relevant images in a database given a text query, which has a wide range of use-cases such as digital libraries [107], e-commerce [260], and multimedia databases [90, 273]. Consequently, there is a growing interest in developing effective models for this task. Current state-of-the-art methods predominantly employ Multimodal Large Language Models (MLLMs) such as BEiT-3 [252] and BLIP [124]. These models generate embeddings for both visual and textual inputs, mapping them into a shared space. The mapping function is usually designed to be injective, facilitating a one-to-one correspondence between an instance and its point in the embedding space. Fine-tuning these MLLMs on smaller image-caption datasets such as MSCOCO [133] and Flickr30K [274] enables the models to achieve high accuracy in text-to-image retrieval tasks.

However, MLLMs-based methods face limitations particularly in the context of real-world use-cases that involve large-scale, diverse, and ambiguous data such as that illustrated in Figure 9.1 and Figure 9.2. First, MLLMs-based methods often ignore efficiency concerns. Their model-based similarity inference methods [84] are computationally demanding, requiring encoding between each query vector and image embedding when ranking. This can result in a computation time of up to 22 hours for a single inference for a large test set [266], limiting their utility in large-scale retrieval applications despite their high accuracy. Second, real-world use-cases often involve complex queries and images with multiple objects [221, 223, 266]. This contrasts sharply with the comprehensive but short captions found in datasets such as MSCOCO [133] and Flickr30K [274]. The nature of this complexity undermines the effectiveness of injective embeddings, which attempt to map diverse meanings/senses to a single point in shared space, which could be an inaccurate weighted geometric mean of all the desirable points [221]. This is particularly problematic in long-text query to image retrieval tasks, where accumulated ambiguities significantly hinder the performance of Multimodal Large Language Models (MLLMs) [266]. Third, injective embeddings struggle with partial text-to-image associations [221]. In a long query, only a subset of sentences may relate to specific regions or aspects of an image, while the rest discuss unrelated subjects. Additionally, a single sentence may describe just a particular region of an image rather than its entirety.

To address these challenges and based on the framework we proposed in Chapter 4, this chapter presents a novel two-stage Coarse-to-Fine Index-shared Retrieval (CFIR) framework, jointly optimizing effectiveness and efficiency in Section 9.2. The CFIR framework is designed based on the proposed multimodal learning framework, MCA (Chapter 4). The first stage is entity-based Ranking (ER) and the ER result is used to construct a shared entity-based image candidates index, as described in Section 9.2.2. ER is designed to be computationally cheap, using pre-computed image embeddings from a cache. By replacing the entire document with a representation comprising its entities as the query, we transform the retrieval task from one query to one target, to multiple queries to multiple targets, accommodating the ambiguity inherent in long documents and images. This transformation makes ER well-suited for use-cases demanding relevance but not exact matching, such as multimedia content creation [44], where a diverse array of images is beneficial for illustrative purposes. Furthermore, ER can be used to filter out the majority of irrelevant candidates prior to the re-ranking stage, thereby reducing the overhead from the more powerful encoder used in the re-ranking stage. The second stage is Summary-based Re-ranking (SR), as shown in Section 9.2.3. By summarizing long documents as queries and using entity-based image candidates from the pre-computed shared index, SR further mitigates ambiguity, making the framework robust against partial text-to-image associations and reducing encoding time. The main contributions of this work are as follows:

1. We introduce the two-stage Coarse-to-Fine Index-shared Retrieval (CFIR) framework to address the effectiveness and efficiency challenges of state-of-the-art MLLM-based approaches in real-world scenarios. The framework includes Entity-based Ranking (ER) and Summary-based Ranking (SR) stages. This supports our thesis statement by demonstrating how structured methodologies can enhance multimodal learning performance and efficiency.

2. The Entity-based Ranking (ER) stage innovates beyond the prevalent one-to-one retrieval paradigm in MLLM-based methods by employing a multiple-queries-to-multiple-targets approach. This enhances ambiguity handling and improves performance by efficiently fil-


A Lengthy Document Query from AToMic
Multi-faceted Corresponding Matching Images
[1]
[3]
[4]
[2]
Short Captions Query from MSCOCO
Exact Matching Image

'page_title': 'Space Jam', 'section_title': 'Live-action', 'context_page_description': "Space Jam is a 1996 American live-action/animated sports comedy film directed by Joe Pytka, with animation sequences directed by Bruce W. Smith and Tony Cervone**[1]**, and written by Leo Benvenuti, Steve Rudnick, Timothy Harris and Herschel Weingrod. The film stars basketball player Michael Jordan**[2]** as a fictional version of himself; Wayne Knight and Theresa Randle appear in supporting roles, while Billy West, Dee Bradley Baker, Kath Soucie, and Danny DeVito headline the voice cast. The film is a fictionalized account of the timeline ... Arsenio Hall**[3]** ..... Beau Walker **[4]**", 'context_section_description': "Some of the film's live-action cast play fictional versions of themselves:\n Michael Jordan as himself Brandon Hammond as Michael Jordan (10 years old)\n Wayne Knight as Stan Podolak, a publicist and assistant who aids Jordan\n Theresa Randle as Juanita Jordan, Jordan's wife\n Bill Murray as himself; ..."

Caption1: Young basketball players run down the court in a game of basketball .
Caption2: The young boys are playing a game of basketball .

Figure 9.1: Comparison of examples from AToMiC and MSCOCO datasets.

tering candidates for re-ranking, aligning with our thesis hypothesis that improved alignment and integration can boost performance.

3. Our Summary-based Re-ranking (SR) stage utilizes summarized queries, enabling the application of more computationally-intensive models to refine the candidate set generated by ER. This step supports our thesis statement by showing how effective resource utilization can enhance the overall efficiency of multimodal learning frameworks.

4. We introduce a novel Decoupling-BEiT-3 encoder optimized for both ER and SR stages, as detailed in Section 9.2.1. This encoder employs a decoupled encoding design for vector-based distance computation, enhancing both training and retrieval efficiency. The use of an entity-based image candidate index and a pre-computed image embedding cache, based on a frozen vision encoder, significantly improves large-scale application performance. This innovation supports our hypothesis by demonstrating the effectiveness of optimized model designs in improving multimodal learning outcomes.

5. CFIR is evaluated on the AToMiC dataset, showing an 11.06% improvement in Recall@1000 and reducing computational times by 68.75% and 99.79% in training and retrieval, respectively, as detailed in Section 9.4. These results validate our thesis statement by providing empirical evidence that our proposed framework enhances both performance and efficiency in real-world applications.

6. The original material in this section has been accepted for presentation at the 2024 International ACM SIGIR Conference on Research and Development in Information Retrieval, a Core ranking A* conference with an h5-index of 103.

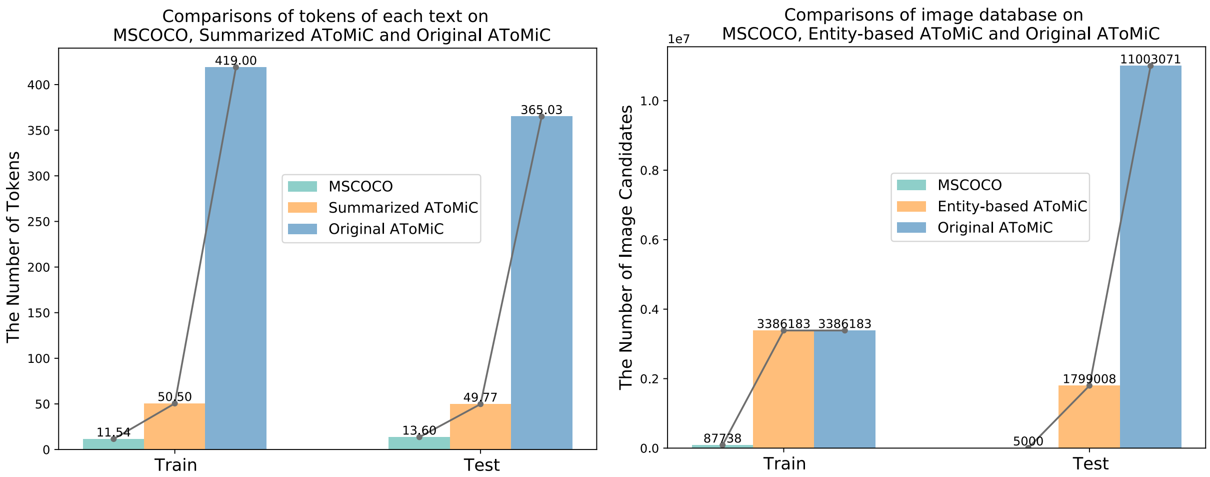


Figure 9.2: The left is a plot of the average text tokens between MSCOCO short sentences, AToMiC long documents and their summaries and the right is the number of training and testing images of MSCOCO and AToMiC.

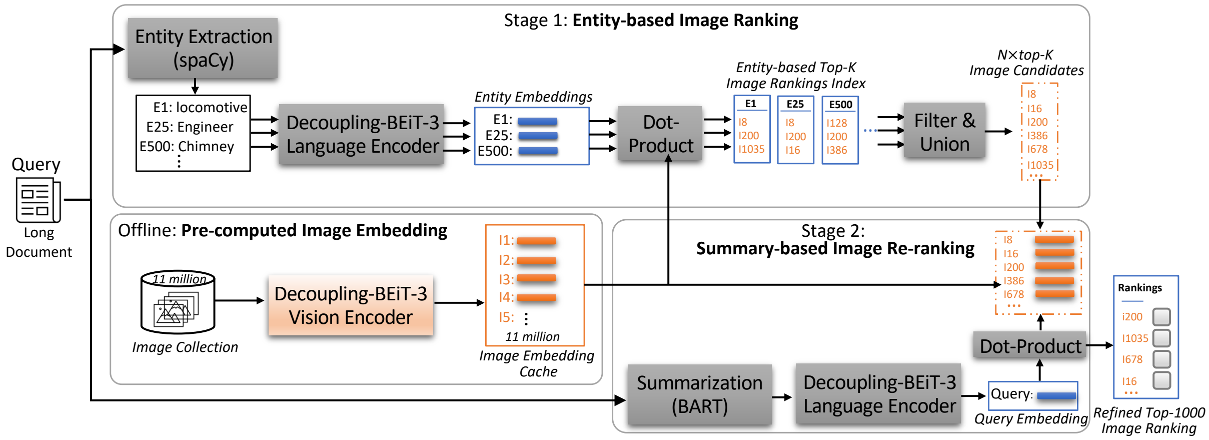


Figure 9.3: The overall architecture of the proposed CFIR for large-scale document-to-image retrieval.

9.2 *CFIR:* Fast and Effective Document-To-Image Retrieval for Large Corpora

To systematically address the challenges inherent in Multimodal Large Language Models (MLLMs) for Large-Scale Long-Text to Image Retrieval (LLIR), we propose a two-stage coarse-to-fine index-shared retrieval (CFIR) framework, as shown in Figure 9.3. The pseudocode for the corresponding training algorithm is shown in Figure 9.4. Moreover, the pseudocode for the corresponding retrieval algorithm (during testing) is shown in Figure 9.5. CFIR is subdivided into two core stages: Entity-based Ranking (ER) and Summary-based Re-ranking (SR). We also introduce a novel Decoupling-BEiT-3 encoder optimized for both ER and SR stages.

```
Require: Long-text set $\mathscr{D}$, Image set $\mathscr{I}$, A pre-trained version of our proposed decoupling-BEiT-3 model, Text-entity extractor spaCy, Text-summary generator BART large model [119].
Ensure: Entity-based image ranking index $\mathscr{E}_{index}$, Image embedding index $\mathscr{V}_{index}$.
    // Stage 0: Pre-computing image embedding index.
    for for each image in $\mathscr{I}$ do do
        Encode the image using our proposed D-BEiT-3 model with image expert to generate embedding $v_i$;
        $\mathscr{V}_{index}$.append($v_{i \to index}$);
    end for
    // Stage 1: Entity-based Ranking (ER).
    for each long-text query in $\mathscr{D}$ do
        Extract entities $\{e_1, \ldots, e_N\}$ of $i$-index long-text query by spaCy;
        Encode $e_j$ using our proposed D-BEiT-3 model with text encoder to generate embedding $t_j$;
        Build a similarity score list $S$;
        for for each image embedding in $\mathscr{V}$ do do
            Compute the similarity (dot-product) between $t_j$ and the selected image embedding; $\mathscr{S}_{index}$.append($t_{j \to index}$);
        end for
        Choose Top-K from $S$ to build $\mathscr{E}_{i \to index}$ = list($[I_1, \ldots, I_K]$);
    end for
    // Stage 2: Summary-Based Re-ranking (SR).
    for each-index long-text query in $\mathscr{D}$ do
        Extract the entities from $i$-index long-text query by spaCy;
        Obtain the corresponding pre-stored Top-K image ranking index from $\mathscr{E}_{index}$ based on the extracted entities;
        Filter & Union repeated candidates ranking index form a candidate set;
        Obtain the corresponding image embedding set $\mathscr{V}_{candidates}$ from $\mathscr{V}_{index}$;
        Summary the $i$-index long-text query by the BART large model;
        Encode the summary by using our proposed D-BEiT-3 with text expert as $\mathbf{q}_{index}$;
        Compute the similarities (dot-product) between the query embedding $\mathbf{q}_{index}$ and coarse-grained image embedding set $\mathscr{V}_{candidates}$;
        return image ranking.
    end for
```

Figure 9.4: CFIR Training Procedure

```
Require: long-text query $q$, A pre-trained version of our proposed decoupling-BEiT-3 model,
    Text-entity extractor spaCy, Text-summary generator BART large model [119]. Entity-
    based image ranking index $\mathscr{E}_{index}$, Image embedding index $\mathscr{V}_{index}$.
    // Summary-Based Re-ranking (SR).
    Extract the entities from the long-text query $q$ by spaCy;
    Obtain the corresponding pre-stored Top-K image ranking index from $\mathscr{E}_{index}$ based on the
        extracted entities;
    Filter & Union repeated candidates ranking index form a candidate set;
    Obtain the corresponding image embedding set $q_{candidates}$ from $\mathscr{V}_{index}$;
    Summary the long-text query $q$ by the BART large model;
    Encode the summary by using our proposed D-BEiT-3 with text expert as $\mathbf{q}_{encoded}$;
    Compute the similarities (dot-product) between the query embedding $\mathbf{q}_{encoded}$ and image
        embedding set $q_{candidates}$;
    return image ranking.
```

Figure 9.5: CFIR Retrieval (testing) Procedure

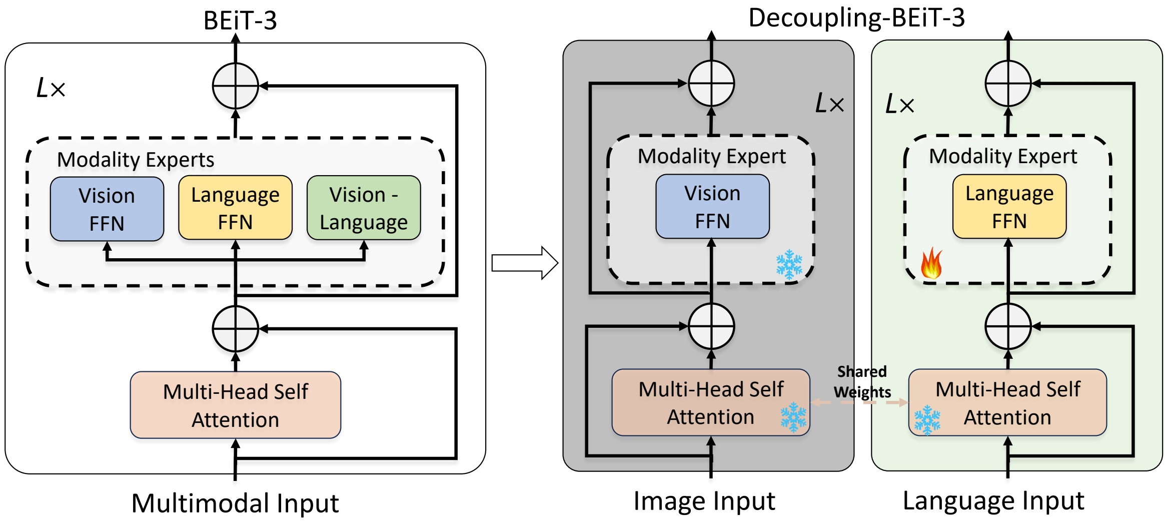


Figure 9.6: The demonstration of differences between the original architecture of BEiT-3 model and our Decoupling-BEiT-3.

9.2.1 The Proposed Decoupling-BEiT-3

The BEiT-3 model is originally constructed as a MultiWay Transformer design [252]. As depicted on the left side of Figure 9.6, the MultiWay Transformer block in BEiT-3 features shared self-attention modules and a pool of feed-forward networks (i.e., modality experts) tailored for different modalities.

To better fit the LLIR task, in this section, we propose a decoupling-BEiT-3 (D-BEiT-3) as the MLLM encoder in our CFIR. Our D-BEiT-3 architecture removes the Vision-Language (VL) expert, as shown on the right side of Figure 9.6. This design is motivated by three primary considerations. First and most importantly, without the VL expert, we decouple the encoding of visual and text input and transition from model-based similarity inference to vector-based distance computation, which is significantly faster. We also index the image vector to further reduce computational cost during both training and testing. Specifically, if we were to use the VL expert in the original BEiT-3 model, it would be necessary in the inference stage to exhaustively pair the query with each database item and then feed these pairs into the BEiT-3 model to predict matching scores. Second, although BEiT-3's design is effective for accurate instance-level alignment between text and images, it is optimized for image descriptions (captions) that are both precise and comprehensive. This specialization is at odds with the multi-faceted (ambiguous), long documents found in the AToMiC dataset, which often describe multiple images and objects. This inherent ambiguity causes the model to underperform, as identified in [266]. By eliminating the vision-language expert, our architecture better suits the less stringent semantic alignment requirements of the LLIR task. Third, our streamlined architecture results in a 30.4% reduction in model parameters compared to the original BEiT-3 model, significantly enhancing both training and inference efficiency.

9.2.2 Entity-based Ranking (ER)

The Entity-based Ranking (ER) stage serves two primary functions. First, it generates a Top-K ranking of images for each unique named entity extracted from long-text queries, thereby mitigating ambiguity and partial associations. This is achieved through a shift from a one-to-one to a multiple-queries-to-multiple-targets retrieval paradigm. Second, ER effectively prunes irrelevant image candidates, paving the way for the subsequent re-ranking. To facilitate this and improve efficiency, we construct an entity-based candidate index that maps each entity to its likely corresponding images, based on the Top-K ranking obtained from the ER stage. Repeated entities across different query documents can be swiftly retrieved from the index, reducing computational costs. Consequently, in the training phase, we construct an entity-based candidate index encompassing all ranking results for entities present in the training samples. If an unknown entity (not included in the training samples) appears during retrieval, it is disregarded. This approach has a negligible impact on performance, given the extensive dataset of over 11 million training samples. Additionally, it significantly boosts efficiency by obviating the necessity to recompute any Entity Retrieval (ER) stage for new entities during test retrieval. This approach allows for the computation of the entity-based ranking to be performed only once, during the training phase. Furthermore, if there is a need to augment system performance by incorporating additional unknown entities into the entity-based candidate index, this can be efficiently achieved offline. This entails appending the ranking results to the index at a later time, rather than conducting this process online during retrieval, which has no impact on the retrieval efficiency.

To accomplish this, as shown in Figure 9.3, we first employ the advanced Natural Language Processing (NLP) library, spaCy, to extract name entities from each long document. SpaCy's robust entity extraction capabilities serve as an effective mechanism for generating entity queries.

Subsequently, we use a pre-trained and frozen D-BEiT-3 to encode these entities, eliminating the need for additional training and thereby enhancing computational efficiency. We then retrieve the Top-K candidate images based on their similarity scores with each entity. These similarity scores are calculated as the dot-product between each entity's embedding and the embeddings of the image candidates, which are retrieved from a pre-computed shared image embedding cache.

9.2.3 Summary-Based Re-ranking (SR)

To achieve precise image matching, we introduce the Summary-based Re-ranking (SR) stage, tailored for LLIR. Contrary to the ER stage, SR focuses exclusively on precise image matching of a document's key information - document summary - to refine the ranking of previously identified entity-based image candidates. Text summaries are generated using BART [119], chosen for its ability to produce concise summaries that capture the document's core semantics. Notably, BART [119] supports a larger maximum input token count of 1024, accommodating the average token number of 419 in AToMiC queries, compared to BERT's [46] 512-token limit. The summary effectively mitigates the semantic ambiguity and partial association problems in long document query, thus improving the retrieval effect.

During training, only the language expert component of our D-BEiT-3 is fine-tuned, leaving the remaining modules frozen. This approach facilitates the construction of a pre-computed shared image embedding cache, striking an optimized balance between training efficiency and retrieval efficacy. The entire training process is optimized by a symmetric cross-entropy loss over the similarity scores between the text representation and image representations.

In inference, we utilize pre-computed image embeddings from a shared cache, negating the need for recalculations in each training epoch. Candidate selection bypasses full-database retrieval, opting for a union subset comprising the top-K entity-queried candidates for each entity in the query. The theoretical candidate set size should be $N \times$Top-K, where N is the number of entities in a query, a count significantly lower than that of the comprehensive image database. For instance, with 10 entities and selecting the Top-10,000 results for each, we have an image candidate pool of 100,000. This size is substantially smaller compared to the 4 million images in the AToMiC base setting or the 11 million images in the larger setting. This strategy reduces computational load and is made possible by the shared entity-based candidates index. Moreover, empirical observations indicate a substantial overlap—approximately 45.6%—between candidate sets for distinct entities, resulting in an actual filtered set size substantially smaller than $N \times$Top-K. Consequently, it is only necessary to calculate the dot-product between the embeddings of filtered image candidates and the summary query embeddings to establish the final ranking. This approach eliminates the need for computations involving all images in the database.

9.3 Experimental Setup

We utilize the Base (H = 768) and Large model (H = 1024) of our streamlined version of BEiT-3 as our encoder in CFIR, denoted as CFIR-B and CFIR-L respectively, where H is the hidden size. Additionally, we include comparisons with two state-of-the-art Multimodal Large Language models: our proposed D-BEiT-3 and OpenCLIP [36]. OpenCLIP is an open-source variant of OpenAI's CLIP, specialized for multi-modal learning with text and images. It enables zero-shot classification and cross-modal retrieval in a shared embedding space without requiring task-specific fine-tuning.

We adhere to the experimental setup for BEiT-3 [1], and fine-tune for the AToMiC dataset. For OpenCLIP we include the results reported in [266]. We conduct training over 30 epochs. For image data augmentation in training, we employ AutoAugment [41]. Throughout all fine-tuning experiments, we choose the Adam optimizer with a learning rate set at 1×10^{-4}, a weight decay of 0.05, and a batch size of 512. We also integrate a dropout rate of 0.1.

Adhering to creator of AToMiC dataset [266] and to ensure a fair comparison, we assess the performance of all methods using established metrics: recall at 1000 (R@1000) and mean reciprocal rank at 10 (MRR@10). Additionally, we report both training and retrieval times to evaluate model efficiency.

9.4 Experiments

This section aims to evaluate the efficacy and efficiency of our proposed CFIR framework in addressing the challenges outlined in Section 2.11.2. Our evaluation encompasses both the base and large settings of the AToMiC dataset, as discussed in Section 2.11.2. Specifically, we answer the following research questions.

- RQ1: What is the Benefit and Cost of Freezing the Image Encoder?
- RQ2: How does CFIR perform compared to state-of-the-art approaches?
- RQ3: How does the CFIR model demonstrate scalability in the context of the larger and more challenging AToMiC Large Setting?

9.4.1 What is the Benefit and Cost of Freezing the Image Encoder? (RQ1)

In this section, we evaluate the computational and performance trade-offs of freezing the image encoder by comparing its performance with that of whole-model fine-tuning. The comparison focuses on two models: OpenCLIP and D-BEiT-3, as detailed in upper block of Table 9.1 for

[1] https://github.com/microsoft/unilm/tree/master/beit3

Table 9.1: Comparisons of experimental results on AToMiC base setting for large-scale document-to-image retrieval. VE and LE indicate the vision encoder and language encoder with (🔥) or without (❄) fine-tuning. # P (M) indicates trainable parameters of the multi-modal encoders. T-t (Hour/epoch) means the training time and R-t (millisecond/query) means the retrieval time for each query. For OpenCLIP runs, - indicates the metric was not reported in [266].

Method	VE	LE	# P	T-t	AToMiC Base Setting		
					R-t	MRR@10	R@1000
OpenCLIP-B	🔥	🔥	197	-	-	0.043	44.68
D-BEiT-3-B (proposed model)	🔥	🔥	155	16	1640.7	0.048	50.65
OpenCLIP-L	🔥	🔥	645	-	-	0.065	54.84
D-BEiT-3-L (proposed model)	🔥	🔥	490	76	2257.5	0.085	57.39
OpenCLIP-B	❄	🔥	110	-	-	0.037	39.66
D-BEiT-3-B (proposed model)	❄	🔥	87	8	363.6	0.042	43.28
CFIR-B (proposed framework)	❄	🔥	87	**5**	**4.2**	**0.052**	**50.72**
OpenCLIP-L	❄	🔥	340	-	-	0.063	50.13
D-BEiT-3-L (proposed model)	❄	🔥	305	53	425.1	0.065	51.36
CFIR-L (proposed framework)	❄	🔥	305	**45**	**4.7**	**0.081**	**55.68**

Table 9.2: Comparisons of experimental results on large setting for large-scale document-to-image retrieval.

Method	VE	LE	# P	T-t	AToMiC Large Setting		
					R-t	MRR@10	R@1000
D-BEiT-3-B (proposed model)	🔥	🔥	155	16	6016.3	0.019	37.11
D-BEiT-3-L (proposed model)	🔥	🔥	490	76	8140.2	0.038	43.37
D-BEiT-3-B (proposed model)	❄	🔥	87	8	1349.1	0.015	36.16
CFIR-B (proposed framework)	❄	🔥	87	**5**	**364.1**	**0.021**	**38.07**
D-BEiT-3-L (proposed model)	❄	🔥	305	53	1458.5	0.026	39.36
CFIR-L (proposed framework)	❄	🔥	305	**45**	**364.6**	**0.030**	**42.51**

AToMiC base setting and Table 9.2 for AToMiC large setting. The primary motivation for freezing the image encoder lies in the creation of an image embedding cache, which substantially mitigates computational overhead on encoding images during both training and retrieval phases.

In the AToMiC base setting, when compared to the fully fine-tuned OpenCLIP large model (OpenCLIP-L-Full), D-BEiT-3 large model with frozen image encoder (D-BEiT-3-L-Frozen) demonstrates a modest drop of 3.48% in Recall@1000, making the performance loss acceptable given that OpenCLIP-L was the previous state-of-the-art model. Moreover, D-BEiT-3-L-Frozen experiences a decrease in performance, with a 6.03% drop in Recall@1000 and a 0.02 reduction in MRR@10, compared to its whole-model fine-tuned version (D-BEiT-3-L-Full). However, D-BEiT-3-L-Frozen is fine-tuned with only 56% of the parameters and achieves a 30% reduction in training time compared to the OpenCLIP-L-Full. Furthermore, the implementation of an image embedding cache leads to a substantial reduction in retrieval time. The query time decreases by 81% for D-BEiT-3-L (from 2257.5 ms to 425.1 ms) and by 77.8% for D-BEiT-3 base model (D-BEiT-3-B) (from 1640.7 ms to 363.6 ms) when transitioning from fully D-BEiT-3-L/B-Full to D-BEiT-3-L/B-Frozen. This leads to a considerable reduction in training time. For instance, D-BEiT-3-L-Frozen shows a decrease of 23 hours per epoch compared to D-BEiT-3-L-Full, while D-BEiT-3-B-Frozen exhibits an 8-hour reduction per epoch relative to D-BEiT-3-B-Full. This is particularly significant as model testing forms an integral part of the training process. Employing a whole-model fine-tuning approach would necessitate approximately 2280 hours (around 95 days) to train the model over 30 epochs on a single GPU. Such a duration is impractical in typical academic experimental environments.

As for the more challenging scenario, AToMiC Large setting which has 11 millions long-text and 11 millions images, D-BEiT-3-L-Frozen demonstrates remarkable efficiency. As for another important aspect in evaluating the efficiency, the retrieval latency, D-BEiT-3-L-Frozen and D-BEiT-3-B-Frozen needs just 17.9% and 22.4% of the retrieval time required by D-BEiT-3-L-Full and D-BEiT-3-B-Full. This significant reduction in retrieval time underscores the enhanced scalability and aptness of these models for large-scale applications.

In summary, we construct an shared image embedding cache and a shared entity-based image ranking index to markedly enhance training and retrieval efficiency, predicated on the freezing of the image encoder. While there is a performance trade-off, the gains in efficiency render the model highly deployable and make it possible for large-scale text-to-image tasks within an academic budget. This result answers our proposed research question 1 that what is the benefit and cost of freezing the image encoder. Consequently, we choose to build our CFIR framework on this premise.

9.4.2 How does CFIR perform compared to state-of-the-art approaches? (RQ2)

This section investigates the improved performance facilitated by CFIR under AToMiC base setting by comparing it with OpenCLIP and the D-BEiT-3 model, mainly in text-only fine-tuning settings, because it is a fair comparison to CFIR. We also compare CFIR with whole-model fine-tuned approaches to demonstrate the performance trade-off.

In the text-only fine-tuning setting, experimental results are illustrated in the bottom block of Table 9.1. For effectiveness, we observe that CFIR outperforms the previous state-of-the-art models OpenCLIP-Frozen and our proposed D-BEiT-3-Frozen under the AToMiC base across all metrics. Notably, CFIR is more effective with smaller encoder sizes. For instance, CFIR-B outperforms OpenCLIP-B-Frozen, achieving a 0.015 higher score in MRR@10 and an 11.06% better result in R@1000. Similarly, CFIR-L surpasses OpenCLIP-L-Frozen with a 0.018 increase in MRR@10 and a 5.55% improvement in R@1000. When compared to D-BEiT-3-B-Frozen, CFIR-B shows a 0.01 improvement in MRR@10 and a significant 7.44% gain in R@1000. Against D-BEiT-3-L-Frozen, CFIR-L leads in both metrics, showing a 0.016 higher score in MRR@10 and a 4.32% advance in R@1000.

In terms of efficiency, compared to prior state-of-the-art MLLM-based methods [36, 252], our approach incurs additional computational time and storage for constructing the entity-based image candidate index and the shared image embedding cache. Specifically, the largest variant of CFIR (CFIR-L) requires an extra 71 GB of storage space and 34 hours of preparation time, which involves building the entity-based image candidate index and the shared image embedding cache. For context, BEiT-3-L requires 2280 hours for a 30-epoch training cycle, the extra 34 hours for CFIR-L's setup is quite minimal which only about 0.14% of the total time BEiT-3-L needs for training. Furthermore, the integration of an index and cache markedly reduces the training duration for CFIR-L. It only requires 45 hours per epoch, culminating in a total of 930 hours for 30 epochs. Owing to the Decoupling-BEiT-3 architectural efficiency and vector-based distance computation, which involves an entity-based image candidates index for filtering and image embedding cache, CFIR significantly streamlines the retrieval process. In the AToMiC base setting, CFIR-B has managed to reduce the retrieval time significantly, from 363.6 milliseconds to mere 4.2 milliseconds, when compared to D-BEiT-3-B-Frozen. For CFIR-L, the reduction in retrieval time is even more remarkable, the time required for CFIR-L to retrieve images is only about 1.1% of the time taken by D-BEiT-3-L-Frozen.

When comparing to full-model fine-tuning approaches, our CFIR still outperforms the previous state-of-the-art model OpenCLIP-Full. CFIR-B gains a 6.04% improvement on Recall@1000 compared to OpenCLIP-B-Full. When compared with the robust whole-model fine-tuned D-BEiT-3-Full, CFIR-B not only slightly surpasses BEiT-3-B-Full in performance but does so with only 56.12% of its parameters, consequently reducing training time by 68.7%. In terms of large models, CFIR-L sustains competitive retrieval performance in comparison to D-BEiT-3-L-

Table 9.3: Ablation studies on AToMic base setting. T-t means training time (Hour/epoch) and R-t means retrieval time (millisecond/query).

Method	Cache	Index	Entity	Summary	Doc	MRR@10	R@1000	T-t	R-t
CFIP-L	-	-	-	-	✓	0.065	51.36	53	425.1
	✓	✓	✓	-	-	0.006	13.15	0	0
	-	-	-	✓	-	0.069	53.91	45	425.1
	✓	✓	✓	-	✓	0.075	54.61	53	13.0
	✓	✓	✓	✓	-	0.081	55.68	45	4.7

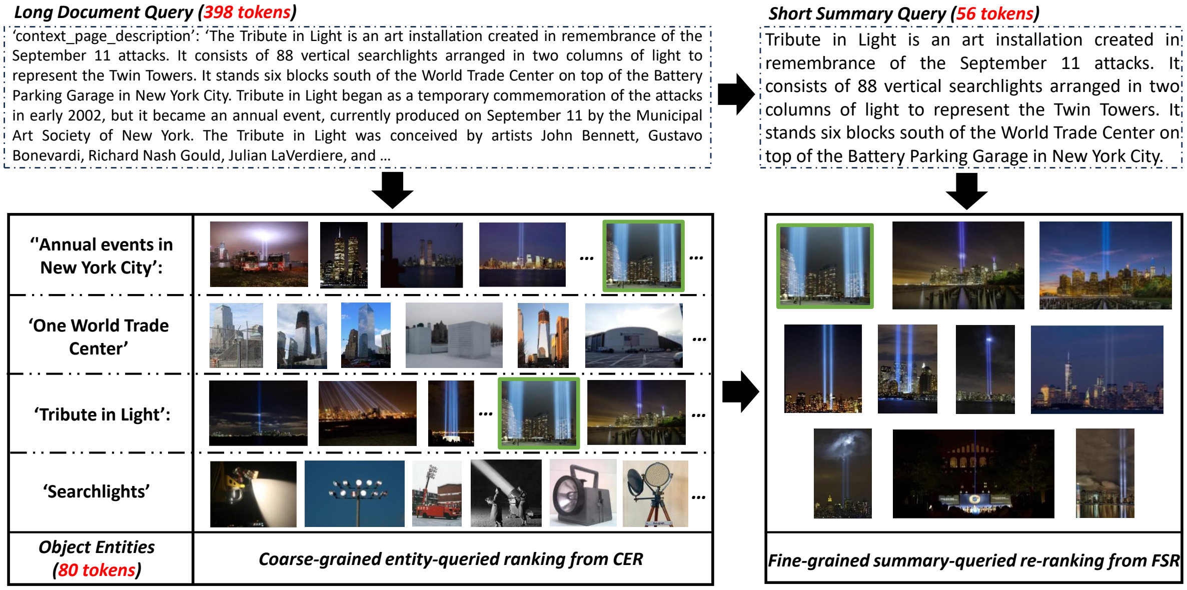
Full, albeit with a minor decline in Recall@1000, while achieving a 40.7% reduction in training time due to the substantial reduction in the length of each document summary. When set against full-model fine-tuning methods, CFIR-B exhibits higher gains in Recall@1000 and compared to D-BEiT-3-B-Full, with 0.07% increase in Recall@1000 in the AToMiC base setting. Similarly, the Recall@1000 performance gap between CFIR-L and D-BEiT-3-L-Full narrows to 1.71%. This result underscores CFIR's better scalability as the candidate set size increases.

To summarize, under AToMiC base setting, CFIR outperforms state-of-the-art models Open-CLIP and D-BEiT-3 in both effectiveness and efficiency in text-only fine-tuning settings. While it incurs modest additional costs, the efficiency gains in training and retrieval time are significant. Even when compared to whole-model fine-tuning approaches, CFIR maintains a competitive performance while requiring fewer parameters and significantly less training and retrieval time, thereby proving its viability for large-scale applications. This result answers our proposed research question 2 that how does CFIR perform compared to state-of-the-art approaches.

9.4.3 CFIR scalability in AToMiC Large Setting (RQ3)

In addition to exploring CFIR's performance in the AToMiC base setting, this section extends the analysis to the more demanding AToMiC large Setting that has three times more images and long-text compare to AToMiC base setting. Here, we also investigates CFIR's enhanced capabilities by comparing it with the D-BEiT-3 model in two scenarios: text-only fine-tuning and whole-model fine-tuning. Combining with our result in AToMiC base setting, we provide a comprehensive comparison between our proposed CFIR framework and previous state-of-the-art methods, and we aim to show CFIR's adaptability and robust performance across diverse fine-tuning scenarios.

The experimental results in the text-only fine-tuning setting, as detailed in the bottom block of Table 9.2, reveal that our proposed CFIR surpasses D-BEiT-3-Frozen in the AToMiC large setting across all evaluated metrics. In particular, CFIR-B shows a notable increase of 0.006 and 1.91% in MRR@10 and R@1000, respectively, when compared to D-BEiT-3-B-Frozen. Similarly, CFIR-L demonstrates a significant lead with improvements of 0.004 and 3.15% in MRR@10 and R@1000, respectively, over D-BEiT-3-L-Frozen. Furthermore, this level of efficiency is consistently observed in the more demanding AToMiC large setting, paralleling results



Long Document Query (398 tokens)

'context_page_description': 'The Tribute in Light is an art installation created in remembrance of the September 11 attacks. It consists of 88 vertical searchlights arranged in two columns of light to represent the Twin Towers. It stands six blocks south of the World Trade Center on top of the Battery Parking Garage in New York City. Tribute in Light began as a temporary commemoration of the attacks in early 2002, but it became an annual event, currently produced on September 11 by the Municipal Art Society of New York. The Tribute in Light was conceived by artists John Bennett, Gustavo Bonevardi, Richard Nash Gould, Julian LaVerdiere, and ...

Short Summary Query (56 tokens)

Tribute in Light is an art installation created in remembrance of the September 11 attacks. It consists of 88 vertical searchlights arranged in two columns of light to represent the Twin Towers. It stands six blocks south of the World Trade Center on top of the Battery Parking Garage in New York City.

"Annual events in New York City':
'One World Trade Center'
'Tribute in Light':
'Searchlights'
Object Entities (80 tokens)
Coarse-grained entity-queried ranking from CER
Fine-grained summary-queried re-ranking from FSR

Figure 9.7: An example of our CFIR for long-text image retrieval.

Table 9.4: The performance impact of varying Top-K in Entity-based Ranking.

Method	Top-K	Retrieval-time	MRR@10	R@1000
CFIP-L	Top-1000	1.5 ms/query	0.033	38.32
	Top-5000	2.9 ms/query	0.051	50.18
	Top-10000	4.7 ms/query	**0.081**	**55.68**
	Top-15000	28.8 ms/query	0.079	55.35

from the AToMiC base setting. In terms of retrieval time, CFIR-B manages to cut down the retrieval time by 985 milliseconds per query compared to D-BEiT-3-B-Frozen. These reductions underscore the effectiveness of CFIR in optimizing the retrieval latency. When set against full-model fine-tuning methods, CFIR-B exhibits higher gains in Recall@1000 and MRR@10 compared to D-BEiT-3-B-Full, with a 0.96%. Similarly, the Recall@1000 performance gap between CFIR-L and D-BEiT-3-L-Full narrows to 0.86%. This result underscores CFIR's better scalability as the candidate set size increases.

In conclusion, within the challenging AToMiC large setting, CFIR demonstrates significantly improvement performance over D-BEiT-3 in terms efficiency in both text-only and whole-model fine-tuning scenarios. These findings underscore CFIR's robust capabilities in enhancing retrieval processes, scaling effective and maintain its power in larger datasets.

9.5 Analysis

This section comprehensively evaluates our CFIR framework through three focused analyses. We dissect CFIR's main components in an ablation study, explore Top-K effects, and offer a qualitative analysis of the model's practical utility.

9.5.1 Ablation Study of CFIR

In this section, we examine the contributions of the Entity-based Ranking (ER) and Summary-based Re-ranking (SR) stages in CFIR, alongside the CFIR-L variant's use of a pre-computed shared index and image embedding cache. We benchmark against state-of-the-art MLLM-based approaches (Row 1, Table 9.3). Using only the SR stage (Row 3, Table 9.3) leads to a significant improvement of 2.55% on R@1000 and reduces training time from 53 to 45 hours per epoch. This indicates that concise summaries enhance both retrieval and training efficiency by reducing document ambiguity.

A comparative analysis (Rows 1 and 4, Table 9.3) shows that integrating ER, a pre-computed shared index, and image embedding cache results in improved retrieval effectiveness. This integration yields an improvement of 0.01 in MRR@10 and 3.25% in R@1000, and reduces the retrieval time from 425.1 millisecond to 13 millisecond per query. The completed CFIR-L model (Row 5, Table 9.3) surpasses all other configurations in both effectiveness and efficiency, requiring significantly less training and retrieval time. Specifically, it cuts down the training time from 53 to 45 hours per epoch and the retrieval time from 425.1 millisecond to 4.7 millisecond per query, while achieving a 0.016 increase in MRR@10 and a 4.32% increase in R@1000 compared to MLLM-based approaches (Row 1, Table 9.3).

In summary, the ER and SR stages in CFIR contribute to both efficient image retrieval and improved performance metrics. Incorporating a pre-computed shared index and image embedding cache further reduces retrieval and training time dramatically.

9.5.2 Performance Impact of Varying Top-K in Entity-based Ranking

In this section, we examine the performance implications of varying the value of Top-K in the Entity-based Ranking (ER) Stage, as detailed in Table 9.4. Our observations indicate that as the value of *k* increases, the re-ranking effectiveness of SR initially experiences a significant boost, only to decrease slightly afterward. Specifically, there is an increase of 0.048 on MRR@10 and 17.36% on R@1000 when Top-K increases from 1000 to 10000. However, this is followed by a slight decrease, from 0.081 to 0.079 in MRR@10 and from 55.68% to 55.35% in R@1000. This phenomenon can be attributed to the increased likelihood of identifying images that are relevant to the query document. However, this also results in the inclusion of additional 'interference' candidates—images that are semantically similar but not identical matches, thereby slightly diminishing effectiveness. Concurrently, the retrieval time shows an upward trend as candidate numbers increase.

To strike a balance between retrieval effectiveness and efficiency, we choose the image candidates for each entity at 10,000 for the SR stage. This configuration necessitates a mere 4.7 millisecond per query in the AToMiC base setting; specifically, it yields scores of 0.081 on MRR@10, and 55.68% on R@1000.

9.5.3 Qualitative Evaluation of CFIR

Figure 9.7 showcases the practical outcomes of applying our CFIR framework within the AToMiC base setting. The figure presents a comprehensive view of the retrieval process, starting with the original long document, which consists of 398 tokens, and proceeding through the crucial stages of entity extraction and summarization. From the long text, 80 entities are extracted, highlighting the essential elements that facilitate the subsequent retrieval tasks. Furthermore, a brief summary of 56 tokens is generated, shorten the document into a concise query. Additionally, it presents the ranked image results yielded by both the ER and SR stages of our framework.

Due to spatial constraints in A4 size, we limit our visualization to four randomly selected entities and their associated top-ranked images, as generated by the first ER stage. This choice in visualization serves to underscore ER's efficacy in not only retrieving a set of relevant images based on individual entities but also in effectively filtering out irrelevant images from a large-scale collection. For instance, the entities "Tribute in Light" and "Annual events in New York City" yield numerous relevant image candidates, including the target image. This approach thereby paves the way for the subsequent SR stage to focus on finding the exact matching image of key document information. As a result, CFIR not only enhances the retrieval accuracy but also speeds up the searching process, ensuring that the most contextually relevant images are brought to the forefront for final selection.

9.6 Conclusions

In this chapter, we address research question 3 proposed in Section 4.3 within the context of cross-modal retrieval. We apply and optimize our proposed multimodal learning framework for cross-modal retrieval tasks. Specifically, we tackle the challenges associated with Large-Scale Long-Text to Image Retrieval (LLIR) by introducing a novel two-stage Coarse-to-Fine Index-Shared Retrieval (CFIR) framework, following the philosophy of deepening intra-modal and inter-modal alignment as stated in our thesis statement (see Section 1.2). Additionally, CFIR is designed to mitigate ambiguity in long documents while optimizing retrieval efficiency by utilizing the efficient transfer learning methods we developed in Chapter 6.

The CFIR framework is modular, comprising two principal stages: Entity-based Ranking (ER) and Summary-based Re-ranking (SR). These stages employ our proposed encoding model, decoupling-BEiT-3, which enables vector-based distance similarity inference and the use of a pre-computed, shared entity-based candidate index and image embedding cache. This significantly improves training and retrieval efficiency.

Our experimental results demonstrate that CFIR outperforms existing state-of-the-art methods in the AToMiC LLIR task, as confirmed by both quantitative and qualitative evaluations. The evidence of increased inter-modality alignment, seen through higher cosine similarity scores between images and text for the same semantics, indicates the effectiveness of our contrastive learning approach. Additionally, the improved performance metrics, such as higher recall rates, showcase the robustness and accuracy of our framework. This comprehensive evaluation supports our thesis statement in Section 1.2, validating our hypothesis that enhanced alignment and integration within multimodal learning frameworks lead to superior performance and efficiency in real-world applications.

These findings have practical implications for various applications that require efficient and effective large-scale image retrieval from long documents. They also confirm the applicability and success of our proposed multimodal learning framework discussed in Chapter 4.

Chapter 10

Conclusions and Future Work

10.1 Contributions and Conclusions

This thesis explores methods to improve the effectiveness and efficiency of multimodal learning, proposing a novel framework, MCA, to integrate these approaches. Specifically, we introduce three key components: the Mixture-of-Modality-Experts (MoME), Contrastive Learning Techniques, and Adapter Methods, each offering distinct advantages. The core of our framework features the MoME component, which enhances computational efficiency and facilitates deeper modality integration through shared transformer block parameters. Contrastive learning methods improve fusion and alignment, enabling better generalization across data types. Adapter-based transfer learning techniques address the practical challenges of efficiently using large models. Through extensive experiments, we have gained insights into the potential and scope of each framework component, demonstrating enhancements in several applications, including crisis response, image-text retrieval, and robotics. The remainder of this section discusses the contributions and conclusions of this thesis in greater detail.

10.1.1 Contributions

In this thesis, we proved that enhancing shallow inter-modal and intra-modal alignment in existing multimodal approaches can improve performance across different tasks by enabling deeper alignment. To address this, we propose a novel multimodal learning framework named MCA, comprising three pivotal components proposed in this thesis—Mixture-of-Modality-Experts (MoME), Contrastive Learning Techniques, and Adapter Methods—will enhance both inter-modal and intra-modal alignment, leading to significantly improved effectiveness and efficiency of multimodal models across a range of tasks. The core of our framework comprises the MoME component, which leverages shared transformer block parameters to enhance computational efficiency and facilitate deeper modality integration, contrastive learning methods to improve fusion and alignment for better generalization across data types, and adapter-based transfer learning techniques to address the practical challenges of using large models efficiently. Our framework MCA outperforms state-of-the-art models, such as LXMERT and VisualBert, in vision-language benchmarks including Visual Question Answering (VQA) and Natural Language for Visual Reasoning (NLVR). Beyond standard benchmarks, our MCA framework undergos extensive testing in real-world applications, focusing on three key downstream tasks: crisis response, image-text retrieval, and robotics. We observe that our enhanced multimodal learning framework will consistently solve real-world problems and exhibit superior performance in effectiveness and efficiency across these diverse domains.

In essence, the main contributions of this thesis are as follows:

- In Chapter 4, we propose a novel multimodal learning framework named MCA, comprising three pivotal components proposed in this thesis—Mixture-of-Modality-Experts (MoME), Contrastive Learning Techniques, and Adapter Methods. These components enhance both inter-modal and intra-modal alignment, leading to significantly improved effectiveness and efficiency of multimodal models across a range of tasks, thereby supporting our thesis statement.

- In Chapter 5, we introduced three methods utilizing contrastive learning to address Research Question 1: How do contrastive learning methods impact modality alignment? These methods enhance the generalization and transferability of Vision Transformers, aiming to produce high-quality visual embeddings and improve inter-modal alignment from various perspectives. In Section 5.2, we present *LaCViT*, a label-aware contrastive fine-tuning framework that significantly improves the Top-1 accuracy of vision transformers across multiple benchmarks. *LaCViT* offers a versatile and comprehensive strategy that greatly enhances the efficacy of transformers for image classification. Our thorough empirical evaluations confirm *LaCViT*'s effectiveness and position it as a viable alternative to the traditional cross-entropy method for fine-tuning pre-trained image classification models. In Section 5.3, we introduce CLCE, a method that integrates label-aware contrastive learning with hard negative mining and cross-entropy (CE) to overcome the limitations of CE and existing contrastive learning techniques. Our empirical data show that CLCE surpasses both traditional CE and earlier contrastive learning methods in both few-shot and transfer learning contexts. Importantly, CLCE is particularly suitable for researchers and developers with access to only commodity GPU hardware, as it achieves effective performance with smaller batch sizes that fit on less powerful GPUs. In Section 5.4, we explore how human labeling errors affect supervised contrastive learning (SCL) differently than they do traditional supervised learning. In response, we develop a new SCL objective, SCL-RHE, which is resistant to real human labeling errors. Our empirical findings indicate that SCL-RHE consistently outperforms traditional cross-entropy approaches, previous SCL objectives, and noise-correcting methods tailored for synthetic noise, in both initial training and transfer learning scenarios. SCL-RHE also stands out for its efficiency—it does not require additional training overhead, unlike methods designed to correct synthetic label noise (shown in Chapter 5).

- In Chapter 6, we introduce the MultiWay-Adapter (MWA) to address Research Question 2: How do parameter-efficient methods improve the efficiency of multimodal learning frameworks? MWA is an effective framework designed for the efficient adaptation of Multimodal Large Language Models (MLLM) to downstream tasks. Addressing the issue of shallow inter-modal alignment in existing methods, MWA employs a dual-component approach, utilizing both the New Knowledge Extractor and the Alignment Enhancer. This strategy enables MWA to not only extract novel information from downstream datasets but also to secure deep inter-modal alignment, as shown in Chapter 6.

- In Chapter 7, we address Research Question 3: How does our proposed multimodal learning framework perform in real-world scenarios? This is explored within the context of crisis response. We investigate multimodal data (vision and language) in Section 7.2, examining the importance of integrating multiple modalities for crisis content categorization. Evaluations with the CrisisMMD dataset show effective automatic labeling, achieving an average of 88.31% F1 performance across two significant tasks (relevance and humanitarian category classification). Our analysis of success and failure cases confirms that deepening intra-modal and inter-modal alignment improves performance in categorizing crisis content on social media. These findings support our thesis statement by demonstrating the practical effectiveness of our proposed multimodal learning framework.

- In Chapter 8, we address Research Question 3: How does our proposed multimodal learning framework perform in real-world scenarios? This is explored within the context of robotic vision. We introduce RoboLLM, based on our MCA framework in Chapter 4, optimized to establish a unified robotic vision pipeline using BEiT-3 as a Multi-Modal Large Language Model backbone encoder. We evaluate our framework on three distinct visual perception tasks: object segmentation, object identification, and defect detection, using the Amazon ARMBench dataset [161]. Our results show that RoboLLM significantly outperforms previous benchmarks across all three challenges. RoboLLM achieves this with inherent knowledge from pretraining on large-scale multimodal data and minimal task-specific heads for each task, significantly increasing performance and mitigating engineering challenges. The modular design of RoboLLM facilitates the incorporation of powerful backbones and task-specific modules, enhancing future performance if required. This versatility and effectiveness demonstrate the applicability and success of our proposed multimodal learning framework, supporting our thesis statement.

- In Chapter 9, we address Research Question 3: How does our proposed multimodal learning framework perform in real-world scenarios? This is explored within the context of cross-modal retrieval. We apply and optimize our proposed multimodal learning framework for the task of cross-modal retrieval. Specifically, we introduce a novel, two-stage Coarse-to-Fine Index-Shared Retrieval (CFIR) framework to address the challenges associated with Large-Scale Long-Text to Image Retrieval (LLIR). CFIR mitigates ambiguity in long documents while optimizing retrieval effectiveness and efficiency through Entity-based Ranking (ER) and Summary-based Re-ranking (SR) stages. Our proposed encoding model, Decoupling-BEiT-3, enhances training and retrieval efficiency through vector-based distance similarity inference and a pre-computed entity-based image candidate index and embedding cache. Experimental results demonstrate that CFIR outperforms existing state-of-the-art methods in the AToMiC LLIR task, showing an 11.06% improvement in Recall@1000 and reducing training and retrieval times by 68.75% and 99.79%, respectively. These findings validate the practical implications of our multimodal learning framework, supporting our thesis statement.

10.1.2 Thesis Conclusions

This section discusses the achievements and conclusions of this thesis, addressing the core hypothesis and validating the thesis statement.

Enhancing Inter-Modal and Intra-Modal Alignment through Contrastive Learning Methods Our experimental results in Chapter 5 confirm that our proposed contrastive learning methods significantly enhance the generalization and transferability of Vision Transformers. These methods result in high-quality visual embeddings and improved inter-modal alignment, demonstrated by better separation in the embedding space. This directly supports our thesis statement by showing that better alignment within multimodal frameworks leads to superior performance.

Efficiency Improvement with the Multi-Way Adapter Our empirical findings in Chapter 6 reveal that adding a mere 2.58% in extra parameters does not result in any statistically significant decline in performance across all tested settings, while reducing fine-tuning time by up to 57%. This efficiency improvement paves the way for future studies on efficient multimodal fine-tuning methods and holds potential for extension into other vision-language tasks. This supports our thesis statement by demonstrating that our framework enhances both effectiveness and efficiency.

Wide Applicability and Enhanced Performance of the Proposed Multimodal Learning Framework in Real-World Scenarios Our evaluations from Chapter 7 to Chapter 9 demonstrated that the framework improves performance across diverse applications, including crisis response, robotic vision, cross-modal retrieval. This versatility and robustness prove the wide applicability of our proposed multimodal learning framework, further validating our thesis statement by showing that the enhanced alignment and integration within multimodal data significantly improve real-world application performance.

10.2 Directions for Future Work

While this thesis has made significant advancements in the field of multimodal learning, several avenues for future research remain open. The following are some promising directions that could further enhance the capabilities and applications of multimodal learning systems:

10.2.1 Advanced Fusion Techniques

Future research can explore more sophisticated methods for fusing multimodal data. Techniques such as dynamic and hierarchical fusion, attention mechanisms, and graph-based models could provide deeper integration and more nuanced interactions between modalities, leading to improved performance.

10.2.2 Scalable and Efficient Architectures

As neural networks continue to grow in size and complexity, developing scalable and efficient architectures will become increasingly important. Future work can focus on creating models that balance performance with computational efficiency, enabling the deployment of multimodal learning systems in resource-constrained environments. Techniques such as model compression, quantization, and efficient transformer designs could play a crucial role in this area.

10.2.3 Transfer Learning and Domain Adaptation

Transfer learning and domain adaptation are essential for applying multimodal models to new tasks and domains with limited labeled data. Future research can investigate ways to improve the transferability of multimodal models and reduce the need for extensive fine-tuning. Approaches such as zero-shot and few-shot learning, meta-learning, and unsupervised domain adaptation could be explored to enhance model generalization across diverse applications.

10.2.4 Robustness and Interpretability

Ensuring the robustness and interpretability of multimodal models is critical for their adoption in real-world applications. Future studies can focus on developing techniques to make these models more resilient to noise, adversarial attacks, and missing data. Additionally, improving the interpretability of multimodal models can help users understand how different modalities contribute to the final predictions, fostering trust and transparency.

10.2.5 Real-Time and Interactive Applications

Real-time and interactive applications of multimodal learning, such as augmented reality (AR), virtual reality (VR), and human-computer interaction (HCI), present unique challenges and opportunities. Research can explore how to optimize multimodal models for low-latency processing and seamless integration with interactive systems. Innovations in this area could lead to more immersive and responsive user experiences.

10.2.6 Ethical and Fair AI

As multimodal learning systems become more pervasive, it is crucial to address ethical and fairness concerns. Future research should focus on developing methods to ensure that these systems are unbiased, equitable, and respect user privacy. This includes creating algorithms that can detect and mitigate biases in multimodal data, as well as establishing guidelines for the ethical deployment of multimodal AI technologies.

10.2.7 Benchmarking and Standardization

The development of standardized benchmarks and evaluation protocols is essential for comparing the performance of different multimodal models. Future efforts should focus on creating comprehensive and diverse benchmark datasets that reflect real-world scenarios. Standardization in evaluation metrics and protocols can facilitate fair comparisons and drive progress in the field.

10.3 Closing Remarks

In this thesis, we have thoroughly explored enhancing shallow inter-modal and intra-modal alignment in existing multimodal approaches to improve performance across different tasks. We developed and applied a novel multimodal learning framework named MCA, comprising three pivotal components: Mixture-of-Modality-Experts (MoME), Contrastive Learning Techniques, and Adapter Methods. Our key findings include significant improvements in both effectiveness and efficiency of multimodal models across a range of tasks, thereby validating our thesis statement.

The significance of this research lies in its potential to revolutionize the field of multimodal learning. By addressing shallow alignment issues, we demonstrated that deeper alignment leads to superior performance, as evidenced by our results in vision-language benchmarks and evaluated downstream tasks. Furthermore, our framework outperformed state-of-the-art models like LXMERT and VisualBert, showcasing its robustness and versatility.

Future research directions include further optimization of the MoME component to enhance computational efficiency, as well as the exploration of novel contrastive learning techniques to improve alignment further. Additionally, extending our adapter methods to more varied and complex downstream tasks could yield even greater performance gains. By building on our findings, future work can refine these methods and potentially lead to new breakthroughs in multimodal learning.

Reflecting on this journey, the process of integrating and building upon prior works from many areas to tackle new tasks and design sound experiments has been both challenging and rewarding. The insights gained during this research have been invaluable and have significantly shaped the direction and outcomes of this thesis.

In conclusion, this thesis has made significant contributions to the field of multimodal learning by enhancing intra-modal and inter-modal alignment, improving computational efficiency, and validating the proposed MCA framework in real-world applications. We observe multiple pieces of evidence for improved intra-modal and inter-modal alignment, along with enhanced performance across various metrics in all four evaluated domains, indicating that our thesis statement holds. These advancements pave the way for future research and development in creating more effective and efficient multimodal systems.

Bibliography

[1] Firoj Alam, Muhammad Imran, and Ferda Ofli. Image4act: Online social media image processing for disaster response. In *Proceedings of the 2017 IEEE/ACM international conference on advances in social networks analysis and mining 2017*, pages 601–604.

[2] Firoj Alam, Ferda Ofli, and Muhammad Imran. Crisismmd: Multimodal twitter datasets from natural disasters. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 12.

[3] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, and Malcolm Reynolds. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736, 2022.

[4] Asha Anoosheh, Torsten Sattler, Radu Timofte, Marc Pollefeys, and Luc Van Gool. Night-to-day image translation for retrieval-based localization. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 5958–5964. IEEE.

[5] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433.

[6] Grigorios Antonellis, Andreas G Gavras, Marios Panagiotou, Bruce L Kutter, Gabriele Guerrini, Andrew C Sander, and Patrick J Fox. Shake table test of large-scale bridge columns supported on rocking shallow foundations. *Journal of Geotechnical and Geoenvironmental Engineering*, 141(5):04015009, 2015.

[7] Eric Arazo, Diego Ortego, Paul Albert, Noel O'Connor, and Kevin McGuinness. Unsupervised label noise modeling and loss correction. In *International conference on machine learning*, pages 312–321. PMLR.

[8] Sanjeev Arora, Hrishikesh Khandeparkar, Mikhail Khodak, Orestis Plevrakis, and Nikunj Saunshi. A theoretical analysis of contrastive unsupervised representation learning. *arXiv preprint arXiv:1902.09229*, 2019.

[9] Sanjeev Arora, Yuanzhi Li, Yingyu Liang, Tengyu Ma, and Andrej Risteski. A latent variable model approach to pmi-based word embeddings. *Transactions of the Association for Computational Linguistics*, 4:385–399, 2016.

[10] Raphael Baena, Lucas Drumetz, and Vincent Gripon. Entropy based feature regularization to improve transferability of deep learning models. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

[11] Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli. Data2vec: A general framework for self-supervised learning in speech, vision and language. In *International Conference on Machine Learning*, pages 1298–1312. PMLR.

[12] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2):423–443, 2018.

[13] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2):423–443, 2018.

[14] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2):423–443, 2018.

[15] Hangbo Bao, Wenhui Wang, Li Dong, Qiang Liu, Owais Khan Mohammed, Kriti Aggarwal, Subhojit Som, Songhao Piao, and Furu Wei. Vlmo: Unified vision-language pre-training with mixture-of-modality-experts. *Advances in Neural Information Processing Systems*, 35:32897–32912, 2022.

[16] Suzanna Becker and Geoffrey E Hinton. Self-organizing neural network that discovers surfaces in random-dot stereograms. *Nature*, 355(6356):161–163, 1992.

[17] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013.

[18] Luca Bertinetto, Joao F Henriques, Philip HS Torr, and Andrea Vedaldi. Meta-learning with differentiable closed-form solvers. *arXiv preprint arXiv:1805.08136*, 2018.

[19] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, and Amanda Askell. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[20] Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: High quality object detection and instance segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 43(5):1483–1498, 2019.

[21] John Canny. A computational approach to edge detection. *IEEE Transactions on pattern analysis and machine intelligence*, (6):679–698, 1986.

[22] Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. *Advances in neural information processing systems*, 32, 2019.

[23] Wenming Cao, Qifan Liu, and Zhiquan He. Review of pavement defect detection methods. *Ieee Access*, 8:14531–14544, 2020.

[24] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in neural information processing systems*, 33:9912–9924, 2020.

[25] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

[26] Hui Chen, Guiguang Ding, Xudong Liu, Zijia Lin, Ji Liu, and Jungong Han. Imram: Iterative matching with recurrent attention memory for cross-modal image-text retrieval. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12655–12663.

[27] Pengfei Chen, Junjie Ye, Guangyong Chen, Jingwei Zhao, and Pheng-Ann Heng. Beyond class-conditional assumption: A primary attempt to combat instance-dependent label noise. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 11442–11450.

[28] Shoufa Chen, Chongjian Ge, Zhan Tong, Jiangliu Wang, Yibing Song, Jue Wang, and Ping Luo. Adaptformer: Adapting vision transformers for scalable visual recognition. *Advances in Neural Information Processing Systems*, 35:16664–16678, 2022.

[29] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR.

[30] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR.

[31] Wenhu Chen, Hexiang Hu, Chitwan Saharia, and William W. Cohen. Re-imagen: Retrieval-augmented text-to-image generator. *ArXiv*, abs/2209.14491, 2022.

[32] Wuyang Chen, Xianzhi Du, Fan Yang, Lucas Beyer, Xiaohua Zhai, Tsung-Yi Lin, Huizhong Chen, Jing Li, Xiaodan Song, and Zhangyang Wang. A simple single-scale vision transformer for object localization and instance segmentation. *arXiv preprint arXiv:2112.09747*, 2021.

[33] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020.

[34] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *European conference on computer vision*, pages 104–120. Springer.

[35] Zetao Chen, Fabiola Maffra, Inkyu Sa, and Margarita Chli. Only look once, mining distinctive landmarks from convnet for visual place recognition. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 9–16. IEEE.

[36] Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2818–2829.

[37] Odysseas S Chlapanis, Georgios Paraskevopoulos, and Alexandros Potamianos. Adapted multimodal bert with layer-wise fusion for sentiment analysis. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

[38] Sumit Chopra, Raia Hadsell, and Yann LeCun. Learning a similarity metric discriminatively, with application to face verification. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, volume 1, pages 539–546. IEEE.

[39] Guanyi Chu, Xiao Wang, Chuan Shi, and Xunqiang Jiang. Cuco: Graph representation with curriculum contrastive learning. In *IJCAI*, pages 2300–2306.

[40] Ching-Yao Chuang, Joshua Robinson, Yen-Chen Lin, Antonio Torralba, and Stefanie Jegelka. Debiased contrastive learning. *Advances in neural information processing systems*, 33:8765–8775, 2020.

[41] Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le. Autoaugment: Learning augmentation strategies from data. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 113–123.

[42] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, volume 1, pages 886–893. Ieee.

[43] Shannon Daly and James A Thom. Mining and classifying image posts on social media to analyse fires. In *ISCRAM*, pages 1–14.

[44] Yashar Deldjoo, Markus Schedl, Paolo Cremonesi, and Gabriella Pasi. Recommender systems leveraging multimedia content. *ACM Computing Surveys (CSUR)*, 53(5):1–38, 2020.

[45] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee.

[46] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[47] Guneet S Dhillon, Pratik Chaudhari, Avinash Ravichandran, and Stefano Soatto. A baseline for few-shot image classification. *arXiv preprint arXiv:1909.02729*, 2019.

[48] Peijian Ding, Davit Soselia, Thomas Armstrong, Jiahao Su, and Furong Huang. Reviving shift equivariance in vision transformers. *arXiv preprint arXiv:2306.07470*, 2023.

[49] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, and Sylvain Gelly. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[50] Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, and Tianhe Yu. Palm-e: An embodied multimodal language model. *arXiv preprint arXiv:2303.03378*, 2023.

[51] Chris Drummond and Robert C Holte. C4. 5, class imbalance, and cost sensitivity: why under-sampling beats over-sampling. In *Workshop on learning from imbalanced datasets II*, volume 11, pages 1–8.

[52] Marthinus C Du Plessis, Gang Niu, and Masashi Sugiyama. Analysis of learning from positive and unlabeled data. *Advances in neural information processing systems*, 27, 2014.

[53] Rafał Dubel, Agata M Wijata, and Jakub Nalepa. On the impact of noisy labels on supervised classification models. In *International Conference on Computational Science*, pages 111–119. Springer.

[54] Alaaeldin El-Nouby, Natalia Neverova, Ivan Laptev, and Hervé Jégou. Training vision transformers for image retrieval. *arXiv preprint arXiv:2102.05644*, 2021.

[55] Charles Elkan and Keith Noto. Learning classifiers from only positive and unlabeled data. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 213–220.

[56] Gamaleldin Elsayed, Dilip Krishnan, Hossein Mobahi, Kevin Regan, and Samy Bengio. Large margin deep networks for classification. *Advances in neural information processing systems*, 31, 2018.

[57] Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. Multiscale vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6824–6835.

[58] Pedro F Felzenszwalb, Ross B Girshick, David McAllester, and Deva Ramanan. Object detection with discriminatively trained part-based models. *IEEE transactions on pattern analysis and machine intelligence*, 32(9):1627–1645, 2009.

[59] Fangxiang Feng, Xiaojie Wang, and Ruifan Li. Cross-modal retrieval with correspondence autoencoder. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 7–16.

[60] Shuo Feng, Huiyu Zhou, and Hongbiao Dong. Using deep neural network with small dataset to predict material defects. *Materials & Design*, 162:300–310, 2019.

[61] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Marc'Aurelio Ranzato, and Tomas Mikolov. Devise: A deep visual-semantic embedding model. *Advances in neural information processing systems*, 26, 2013.

[62] Isaac Chun-Hai Fung, Jingjing Yin, Keisha D Pressley, Carmen H Duke, Chen Mo, Hai Liang, King-Wa Fu, Zion Tsz Ho Tse, and Su-I Hou. Pedagogical demonstration of twitter data analysis: A case study of world aids day, 2014. *Data*, 4(2):84, 2019.

[63] Peng Gao, Zhengkai Jiang, Haoxuan You, Pan Lu, Steven CH Hoi, Xiaogang Wang, and Hongsheng Li. Dynamic fusion with intra-and inter-modality attention flow for visual question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6639–6648.

[64] Weifeng Ge. Deep metric learning with hierarchical triplet loss. In *Proceedings of the European conference on computer vision (ECCV)*, pages 269–285.

[65] Xuri Ge, Fuhai Chen, Joemon M Jose, Zhilong Ji, Zhongqin Wu, and Xiao Liu. Structured multi-modal feature embedding and alignment for image-sentence retrieval. In *Proceedings of the 29th ACM international conference on multimedia*, pages 5185–5193.

[66] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587.

[67] Tiantian Gong, Junsheng Wang, and Liyan Zhang. Rethink pair-wise self-supervised cross-modal retrieval from a contrastive learning perspective. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

[68] Albert Gordo, Jon Almazán, Jerome Revaud, and Diane Larlus. Deep image retrieval: Learning global representations for image search. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VI 14*, pages 241–257. Springer.

[69] Palash Goyal, Sumit Pandey, and Karan Jain. Deep learning for natural language processing. *New York: Apress*, 2018.

[70] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913.

[71] Benjamin Graham, Alaaeldin El-Nouby, Hugo Touvron, Pierre Stock, Armand Joulin, Hervé Jégou, and Matthijs Douze. Levit: a vision transformer in convnet's clothing for faster inference. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 12259–12269.

[72] Gregory Griffin, Alex Holub, and Pietro Perona. Caltech-256 object category dataset. 2007.

[73] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, and Mohammad Gheshlaghi Azar. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020.

[74] Beliz Gunel, Jingfei Du, Alexis Conneau, and Ves Stoyanov. Supervised contrastive learning for pre-trained language model fine-tuning. *arXiv preprint arXiv:2011.01403*, 2020.

[75] Vishal Gupta and Gurpreet S Lehal. A survey of text mining techniques and applications. *Journal of emerging technologies in web intelligence*, 1(1):60–76, 2009.

[76] Michael Gutmann and Aapo Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 297–304. JMLR Workshop and Conference Proceedings.

[77] Bo Han, Quanming Yao, Tongliang Liu, Gang Niu, Ivor W Tsang, James T Kwok, and Masashi Sugiyama. A survey of label-noise representation learning: Past, present and future. *arXiv preprint arXiv:2011.04406*, 2020.

[78] Jiaming Han, Kaixiong Gong, Yiyuan Zhang, Jiaqi Wang, Kaipeng Zhang, Dahua Lin, Yu Qiao, Peng Gao, and Xiangyu Yue. Onellm: One framework to align all modalities with language. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26584–26595, 2024.

[79] Ben Harwood, Vijay Kumar BG, Gustavo Carneiro, Ian Reid, and Tom Drummond. Smart mining for deep metric learning. In *Proceedings of the IEEE international conference on computer vision*, pages 2821–2829.

[80] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009.

[81] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

[82] Markus Hiller, Rongkai Ma, Mehrtash Harandi, and Tom Drummond. Rethinking generalization in few-shot classification. *Advances in Neural Information Processing Systems*, 35:3582–3595, 2022.

[83] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[84] Weixiang Hong, Kaixiang Ji, Jiajia Liu, Jian Wang, Jingdong Chen, and Wei Chu. Gilbert: Generative vision-language pre-training for image-text retrieval. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1379–1388.

[85] Michael E Houle. Local intrinsic dimensionality i: an extreme-value-theoretic foundation for similarity applications. In *Similarity Search and Applications: 10th International Conference, SISAP 2017, Munich, Germany, October 4-6, 2017, Proceedings 10*, pages 64–79. Springer.

[86] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *International Conference on Machine Learning*, pages 2790–2799. PMLR.

[87] Bozhen Hu, Bin Gao, Wai Lok Woo, Lingfeng Ruan, Jikun Jin, Yang Yang, and Yongjie Yu. A lightweight spatial and temporal multi-feature fusion network for defect detection. *IEEE Transactions on Image Processing*, 30:472–486, 2020.

[88] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

[89] Wei Hu, Zhiyuan Li, and Dingli Yu. Simple and effective regularization methods for training on noisily labeled data with generalization guarantee. *arXiv preprint arXiv:1905.11368*, 2019.

[90] Xuming Hu, Zhijiang Guo, Junzhe Chen, Lijie Wen, and Philip S Yu. Mr2: A benchmark for multimodal retrieval-augmented rumor detection in social media. In *Proceedings of the 46th international ACM SIGIR conference on research and development in information retrieval*, pages 2901–2912.

[91] Yuqing Hu, Stéphane Pateux, and Vincent Gripon. Adaptive dimension reduction and variational inference for transductive few-shot classification. In *International Conference on Artificial Intelligence and Statistics*, pages 5899–5917. PMLR.

[92] Chen Huang, Yining Li, Chen Change Loy, and Xiaoou Tang. Learning deep representation for imbalanced classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5375–5384.

[93] Yan Huang, Qi Wu, Chunfeng Song, and Liang Wang. Learning semantic concepts and order for image and sentence matching. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6163–6171.

[94] Zhizhong Huang, Junping Zhang, and Hongming Shan. Twin contrastive learning with noisy labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11661–11670.

[95] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709.

[96] Muhammad Imran, Carlos Castillo, Fernando Diaz, and Sarah Vieweg. Processing social media messages in mass emergency: A survey. *ACM Computing Surveys (CSUR)*, 47(4):1–38, 2015.

[97] Nathalie Japkowicz and Shaju Stephen. The class imbalance problem: A systematic study. *Intelligent data analysis*, 6(5):429–449, 2002.

[98] Piyasak Jeatrakul, Kok Wai Wong, and Chun Che Fung. Classification of imbalanced data by combining the complementary neural network and smote algorithm. In *Neural Information Processing. Models and Applications: 17th International Conference, ICONIP 2010, Sydney, Australia, November 22-25, 2010, Proceedings, Part II 17*, pages 152–159. Springer.

[99] Simon Jenni and Paolo Favaro. Deep bilevel learning. In *Proceedings of the European conference on computer vision (ECCV)*, pages 618–633.

[100] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR.

[101] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. *arXiv preprint arXiv:2102.05918*, 2021.

[102] Ming Jiang, Qiuyuan Huang, Lei Zhang, Xin Wang, Pengchuan Zhang, Zhe Gan, Jana Diesner, and Jianfeng Gao. Tiger: Text-to-image grounding for image caption evaluation. *arXiv preprint arXiv:1909.02050*, 2019.

[103] Ruijie Jiang, Thuan Nguyen, Prakash Ishwar, and Shuchin Aeron. Supervised contrastive learning with hard negative samples. *arXiv preprint arXiv:2209.00078*, 2022.

[104] Yu Jiang, Vivek Natarajan, Xinlei Chen, Marcus Rohrbach, Dhruv Batra, and Devi Parikh. Pythia v0. 1: the winning entry to the vqa challenge 2018. *arXiv preprint arXiv:1807.09956*, 2018.

[105] Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*, 2016.

[106] Yannis Kalantidis, Mert Bulent Sariyildiz, Noe Pion, Philippe Weinzaepfel, and Diane Larlus. Hard negative mixing for contrastive learning. *Advances in Neural Information Processing Systems*, 33:21798–21809, 2020.

[107] Lyndon S Kennedy, Apostol Natsev, and Shih-Fu Chang. Automatic discovery of query-class-dependent models for multimodal search. In *Proceedings of the 13th annual ACM international conference on multimedia*, pages 882–891.

[108] Samuel Kessler, Bethan Thomas, and Salah Karout. An adapter based pre-training for efficient and scalable self-supervised speech representation learning. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3179–3183. IEEE.

[109] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673, 2020.

[110] Erin Hea-Jin Kim, Yoo Kyung Jeong, Yuyoung Kim, Keun Young Kang, and Min Song. Topic-based content and sentiment analysis of ebola virus on twitter and in the news. *Journal of Information Science*, 42(6):763–781, 2016.

[111] Bartosz Krawczyk. Learning from imbalanced data: open challenges and future directions. *Progress in Artificial Intelligence*, 5(4):221–232, 2016.

[112] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. 2009.

[113] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105, 2012.

[114] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

[115] Shamanth Kumar, Geoffrey Barbier, Mohammad Ali Abbasi, and Huan Liu. Tweettracker: An analysis tool for humanitarian and disaster relief. In *Fifth international AAAI conference on weblogs and social media*.

[116] Shamanth Kumar, Fred Morstatter, Reza Zafarani, and Huan Liu. Whom should i follow? identifying relevant users during crises. In *Proceedings of the 24th ACM conference on hypertext and social media*, pages 139–147.

[117] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

[118] Yann LeCun, Bernhard Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne Hubbard, and Lawrence D Jackel. Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1(4):541–551, 1989.

[119] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*, 2019.

[120] Dongdong Li, Zhigang Wang, Jian Wang, Xinyu Zhang, Errui Ding, Jingdong Wang, and Zhaoxiang Zhang. Self-guided hard negative generation for unsupervised person re-identification. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence*.

[121] Feng Li and QingGang Xi. Defectnet: Toward fast and effective defect detection. *IEEE Transactions on Instrumentation and Measurement*, 70:1–9, 2021.

[122] Gen Li, Nan Duan, Yuejian Fang, Ming Gong, and Daxin Jiang. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 11336–11344.

[123] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023.

[124] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR.

[125] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705, 2021.

[126] Kunpeng Li, Yulun Zhang, Kai Li, Yuanyuan Li, and Yun Fu. Visual semantic reasoning for image-text matching. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4654–4662.

[127] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019.

[128] Shikun Li, Xiaobo Xia, Shiming Ge, and Tongliang Liu. Selective-supervised contrastive learning with noisy labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 316–325.

[129] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, and Furu Wei. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX 16*, pages 121–137. Springer.

[130] Yanghao Li, Hanzi Mao, Ross Girshick, and Kaiming He. Exploring plain vision transformer backbones for object detection. In *European Conference on Computer Vision*, pages 280–296. Springer.

[131] Bingqian Lin, Yi Zhu, Zicong Chen, Xiwen Liang, Jianzhuang Liu, and Xiaodan Liang. Adapt: Vision-language navigation with modality-aligned action prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15396–15406, 2022.

[132] Dengtian Lin, Liqiang Jing, Xuemeng Song, Meng Liu, Teng Sun, and Liqiang Nie. Adapting generative pretrained language model for open-domain multimodal sentence summarization. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 195–204.

[133] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer.

[134] Jie Liu, Yixuan Liu, Xue Han, Chao Deng, and Junlan Feng. Escl: Equivariant self-contrastive learning for sentence representations. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

[135] Junhao Liu, Min Yang, Chengming Li, and Ruifeng Xu. Improving cross-modal image-text retrieval with teacher-student learning. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(8):3242–3253, 2020.

[136] Weiyang Liu, Yandong Wen, Zhiding Yu, and Meng Yang. Large-margin softmax loss for convolutional neural networks. *arXiv preprint arXiv:1612.02295*, 2016.

[137] Yahui Liu, Enver Sangineto, Wei Bi, Nicu Sebe, Bruno Lepri, and Marco Nadai. Efficient training of visual transformers with small datasets. *Advances in Neural Information Processing Systems*, 34:23818–23830, 2021.

[138] Yang Liu, Hao Cheng, and Kun Zhang. Identifiability of label noise transition matrix. In *International Conference on Machine Learning*, pages 21475–21496. PMLR.

[139] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[140] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022.

[141] Siqu Long, Soyeon Caren Han, Xiaojun Wan, and Josiah Poon. Gradual: Graph-based dual-modal representation for image-text matching. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3459–3468.

[142] Zijun Long, George Killick, Richard McCreadie, and Gerardo Aragon Camarasa. Multiway-adapater: Adapting large-scale multi-modal models for scalable image-text retrieval. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2024)*.

[143] Zijun Long, George Killick, Richard McCreadie, and Gerardo Aragon Camarasa. Robollm: Robotic vision tasks grounded on multimodal large language models. In *IEEE International Conference on Robotics and Automation (ICRA 2024)*.

[144] Zijun Long, George Killick, Richard McCreadie, Gerardo Aragon Camarasa, and Zaiqiao Meng. When hard negative sampling meets supervised contrastive learning. In *arXiv preprint arXiv:2308.14893*.

[145] Zijun Long, George Killick, Lipeng Zhuang, Richard McCreadie, Gerardo Aragon Camarasa, and Paul Henderson. Elucidating and overcoming the challenges of label noise in supervised contrastive learning. In *arXiv preprint arXiv:2311.16481*.

[146] Zijun Long and Richard McCreadie. Automated crisis content categorization for covid-19 tweet streams. In *18th International Conference on Information Systems for Crisis Response and Management*, pages 667–678.

[147] Zijun Long and Richard McCreadie. Is multi-modal data key for crisis content categorization on social media? In *19th International Conference on Information Systems for Crisis Response and Management (ISCRAM 2022)*.

[148] Zijun Long, Richard McCreadie, Gerardo Aragon Camarasa, and Zaiqiao Meng. Lacvit: A label-aware contrastive fine-tuning framework for vision transformers. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2024)*.

[149] Zijun Long, Richard McCreadie, and Muhammad Imran. Crisisvit: A robust vision transformer for crisis image classification. In *20th International Conference on Information Systems for Crisis Response and Management (ISCRAM 2023)*.

[150] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60:91–110, 2004.

[151] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32, 2019.

[152] Qiwu Luo, Xiaoxin Fang, Li Liu, Chunhua Yang, and Yichuang Sun. Automated visual defect detection for flat steel surface: A survey. *IEEE Transactions on Instrumentation and Measurement*, 69(3):626–644, 2020.

[153] Tomasz Maciejewski and Jerzy Stefanowski. Local neighbourhood extension of smote for mining imbalanced data. In *2011 IEEE symposium on computational intelligence and data mining (CIDM)*, pages 104–111. IEEE.

[154] David Mayo, Jesse Cummings, Xinyu Lin, Dan Gutfreund, Boris Katz, and Andrei Barbu. How hard are computer vision datasets? calibrating dataset difficulty to viewing time. *Advances in Neural Information Processing Systems*, 36, 2024.

[155] Richard McCreadie, Cody Buntain, and Ian Soboroff. Trec incident streams: Finding actionable information on social media. 2019.

[156] Richard McCreadie, Cody Buntain, and Ian Soboroff. Incident streams 2019: Actionable insights and how to find them. 2020.

[157] Iaroslav Melekhov, Juho Kannala, and Esa Rahtu. Siamese network features for image matching. In *2016 23rd international conference on pattern recognition (ICPR)*, pages 378–383. IEEE.

[158] Long Short-Term Memory. Long short-term memory. *Neural computation*, 9(8):1735–1780, 2010.

[159] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[160] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26, 2013.

[161] Chaitanya Mitash, Fan Wang, Shiyang Lu, Vikedo Terhuja, Tyler Garaas, Felipe Polido, and Manikantan Nambi. Armbench: An object-centric benchmark dataset for robotic manipulation. *arXiv preprint arXiv:2303.16382*, 2023.

[162] Marius Mosbach, Maksym Andriushchenko, and Dietrich Klakow. On the stability of fine-tuning bert: Misconceptions, explanations, and strong baselines. *arXiv preprint arXiv:2006.04884*, 2020.

[163] Jiaqi Mu, Suma Bhat, and Pramod Viswanath. All-but-the-top: Simple and effective postprocessing for word representations. *arXiv preprint arXiv:1702.01417*, 2017.

[164] Martin Müller, Marcel Salathé, and Per E Kummervold. Covid-twitter-bert: A natural language processing model to analyse covid-19 content on twitter. *Frontiers in Artificial Intelligence*, 6:1023281, 2023.

[165] Ruchit Nagar, Qingyu Yuan, Clark C Freifeld, Mauricio Santillana, Aaron Nojima, Rumi Chunara, and John S Brownstein. A case study of the new york city 2012-2013 influenza season with daily geocoded twitter data from temporal and spatiotemporal perspectives. *Journal of medical Internet research*, 16(10):e3416, 2014.

[166] Kamil Nar, Orhan Ocal, S Shankar Sastry, and Kannan Ramchandran. Cross-entropy loss and low-rank features have responsibility for adversarial examples. *arXiv preprint arXiv:1901.08360*, 2019.

[167] Dat T Nguyen, Ferda Ofli, Muhammad Imran, and Prasenjit Mitra. Damage assessment from social media imagery data during disasters. In *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017*, pages 569–576.

[168] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian conference on computer vision, graphics & image processing*, pages 722–729. IEEE.

[169] Curtis G Northcutt, Anish Athalye, and Jonas Mueller. Pervasive label errors in test sets destabilize machine learning benchmarks. *arXiv preprint arXiv:2103.14749*, 2021.

[170] Ferda Ofli, Firoj Alam, and Muhammad Imran. Analysis of social media data using multimodal deep learning for disaster response. *arXiv preprint arXiv:2004.11838*, 2020.

[171] Hyun Oh Song, Yu Xiang, Stefanie Jegelka, and Silvio Savarese. Deep metric learning via lifted structured feature embedding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4004–4012.

[172] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[173] OpenAI. Chatgpt: Optimizing language models for dialogue, 2022. Accessed: 2024-05-23.

[174] OpenAI. Gpt-4 technical report, 2023. Accessed: 2024-05-23.

[175] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, and Alaaeldin El-Nouby. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

[176] Vicente Ordonez, Girish Kulkarni, and Tamara Berg. Im2text: Describing images using 1 million captioned photographs. *Advances in neural information processing systems*, 24, 2011.

[177] Boris Oreshkin, Pau Rodríguez López, and Alexandre Lacoste. Tadam: Task dependent adaptive metric for improved few-shot learning. *Advances in neural information processing systems*, 31, 2018.

[178] Diego Ortego, Eric Arazo, Paul Albert, Noel E O'Connor, and Kevin McGuinness. Multi-objective interpolation training for robustness to label noise. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6606–6615.

[179] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3498–3505. IEEE.

[180] Niki Parmar, Ashish Vaswani, Jakob Uszkoreit, Lukasz Kaiser, Noam Shazeer, Alexander Ku, and Dustin Tran. Image transformer. In *International conference on machine learning*, pages 4055–4064. PMLR.

[181] Junyi Peng, Themos Stafylakis, Rongzhi Gu, Oldřich Plchot, Ladislav Mošner, Lukáš Burget, and Jan Černocký. Parameter-efficient transfer learning of pre-trained transformer models for speaker verification using adapters. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

[182] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

[183] Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*, 2018.

[184] Robin Peters and João Porto de Albuquerque. Investigating images as indicators for relevant social media messages in disaster management. In *ISCRAM*.

[185] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649.

[186] Hemant Purohit, Carlos Castillo, Muhammad Imran, and Rahul Pandey. Social-eoc: Serviceability model to rank social media requests for emergency operation centers. In *2018 IEEE/ACM international conference on advances in social networks analysis and mining (ASONAM)*, pages 119–126. IEEE.

[187] Leigang Qu, Meng Liu, Wenjie Wang, Zhedong Zheng, Liqiang Nie, and Tat-Seng Chua. Learnable pillar-based re-ranking for image-text retrieval. *arXiv preprint arXiv:2304.12570*, 2023.

[188] Filip Radenović, Giorgos Tolias, and Ondřej Chum. Fine-tuning cnn image retrieval with no human annotation. *IEEE transactions on pattern analysis and machine intelligence*, 41(7):1655–1668, 2018.

[189] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, and Jack Clark. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

[190] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Saurabh Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2021.

[191] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. 2018.

[192] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[193] Valentin Radu, Catherine Tong, Sourav Bhattacharya, Nicholas D Lane, Cecilia Mascolo, Mahesh K Marina, and Fahim Kawsar. Multimodal deep learning for activity and context recognition. *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies*, 1(4):1–27, 2018.

[194] Abigail Rai and Samarjeet Borah. Study of various methods for tokenization. In *Applications of Internet of Things: Proceedings of ICCCIOT 2020*, pages 193–200. Springer.

[195] Aqsa Rasheed, Bushra Zafar, Amina Rasheed, Nouman Ali, Muhammad Sajid, Saadat Hanif Dar, Usman Habib, Tehmina Shehryar, and Muhammad Tariq Mahmood. Fabric defect detection using computer vision techniques: a comprehensive review. *Mathematical Problems in Engineering*, 2020:1–24, 2020.

[196] Sachin Ravi and Hugo Larochelle. Optimization as a model for few-shot learning. In *International conference on learning representations*.

[197] Sylvestre-Alvise Rebuffi, Hakan Bilen, and Andrea Vedaldi. Learning multiple visual domains with residual adapters. *Advances in neural information processing systems*, 30, 2017.

[198] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788.

[199] Mengye Ren, Eleni Triantafillou, Sachin Ravi, Jake Snell, Kevin Swersky, Joshua B Tenenbaum, Hugo Larochelle, and Richard S Zemel. Meta-learning for semi-supervised few-shot classification. *arXiv preprint arXiv:1803.00676*, 2018.

[200] Mengye Ren, Wenyuan Zeng, Bin Yang, and Raquel Urtasun. Learning to reweight examples for robust deep learning. In *International conference on machine learning*, pages 4334–4343. PMLR.

[201] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015.

[202] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. Bpr: Bayesian personalized ranking from implicit feedback. *arXiv preprint arXiv:1205.2618*, 2012.

[203] Joshua Robinson, Ching-Yao Chuang, Suvrit Sra, and Stefanie Jegelka. A mathematical theory of communication. *arXiv preprint arXiv:2010.04592*, 2020.

[204] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, and Michael Bernstein. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015.

[205] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

[206] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823.

[207] Claude Elwood Shannon. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423, 1948.

[208] Jie Shao, Zhicheng Zhao, and Fei Su. Two-stage deep learning for supervised cross-modal retrieval. *Multimedia Tools and Applications*, 78:16615–16631, 2019.

[209] Ali Sharif Razavian, Hossein Azizpour, Josephine Sullivan, and Stefan Carlsson. Cnn features off-the-shelf: an astounding baseline for recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 806–813.

[210] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565.

[211] Shivam Sharma and Cody Buntain. Improving classification of crisis-related social media content via text augmentation and image analysis. In *TREC*.

[212] Himanshu Shekhar and Shankar Setty. Disaster analysis through tweets. In *2015 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, pages 1719–1723. IEEE.

[213] Abhinav Shrivastava, Abhinav Gupta, and Ross Girshick. Training region-based object detectors with online hard example mining. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 761–769.

[214] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[215] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[216] Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. Flava: A foundational language and vision alignment model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15638–15650.

[217] Kihyuk Sohn. Improved deep metric learning with multi-class n-pair loss objective. *Advances in neural information processing systems*, 29, 2016.

[218] Chull Hwan Song, Jooyoung Yoon, Shunghyun Choi, and Yannis Avrithis. Boosting vision transformers for image retrieval. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 107–117.

[219] Hwanjun Song, Minseok Kim, Dongmin Park, Yooju Shin, and Jae-Gil Lee. Learning from noisy labels with deep neural networks: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

[220] Shezheng Song, Xiaopeng Li, Shasha Li, Shan Zhao, Jie Yu, Jun Ma, Xiaoguang Mao, and Weimin Zhang. How to bridge the gap between modalities: A comprehensive survey on multimodal large language model. *arXiv preprint arXiv:2311.07594*, 2023.

[221] Yale Song and Mohammad Soleymani. Polysemous visual-semantic embedding for cross-modal retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1979–1988.

[222] Zhao Song, Ke Yang, Naiyang Guan, Junjie Zhu, Peng Qiao, and Qingyong Hu. Vppt: Visual pre-trained prompt tuning framework for few-shot image classification. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

[223] Krishna Srinivasan, Karthik Raman, Jiecao Chen, Michael Bendersky, and Marc Najork. Wit: Wikipedia-based image text dataset for multimodal multilingual machine learning. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2443–2449.

[224] Kevin Stowe, Michael Paul, Martha Palmer, Leysia Palen, and Kenneth M Anderson. Identifying and categorizing disaster-related tweets. In *Proceedings of The fourth international workshop on natural language processing for social media*, pages 1–6.

[225] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. Vl-bert: Pre-training of generic visual-linguistic representations. *arXiv preprint arXiv:1908.08530*, 2019.

[226] Yumin Suh, Bohyung Han, Wonsik Kim, and Kyoung Mu Lee. Stochastic class-based hard example mining for deep metric learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7251–7259.

[227] Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. A corpus for reasoning about natural language grounded in photographs. *arXiv preprint arXiv:1811.00491*, 2018.

[228] Yi-Lin Sung, Jaemin Cho, and Mohit Bansal. Vl-adapter: Parameter-efficient transfer learning for vision-and-language tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5227–5237.

[229] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826.

[230] Domen Tabernik, Samo Šela, Jure Skvarč, and Danijel Skočaj. Segmentation-based deep-learning approach for surface-defect detection. *Journal of Intelligent Manufacturing*, 31(3):759–776, 2020.

[231] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*, 2019.

[232] Qingyu Tan, Ruidan He, Lidong Bing, and Hwee Tou Ng. Domain generalization for text classification with memory-based supervised contrastive learning. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6916–6926.

[233] Bethan Thomas, Samuel Kessler, and Salah Karout. Efficient adapter transfer of self-supervised speech models for automatic speech recognition. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7102–7106. IEEE.

[234] Jialin Tian, Kai Wang, Xing Xu, Zuo Cao, Fumin Shen, and Heng Tao Shen. Multimodal disentanglement variational autoencoders for zero-shot cross-modal retrieval. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 960–969.

[235] Yonglong Tian, Yue Wang, Dilip Krishnan, Joshua B Tenenbaum, and Phillip Isola. Rethinking few-shot image classification: a good embedding is all you need? In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16*, pages 266–282. Springer.

[236] Hien To, Sumeet Agrawal, Seon Ho Kim, and Cyrus Shahabi. On identifying disaster-related tweets: Matching-based or learning-based? In *2017 IEEE third international conference on multimedia big data (BigMM)*, pages 330–337. IEEE.

[237] Giorgos Tolias, Ronan Sicre, and Hervé Jégou. Particular object retrieval with integral max-pooling of cnn activations. *arXiv preprint arXiv:1511.05879*, 2015.

[238] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, pages 10347–10357. PMLR.

[239] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[240] Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8769–8778.

[241] Steven Vander Eeckt and Hugo Van Hamme. Using adapters to overcome catastrophic forgetting in end-to-end automatic speech recognition. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

[242] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[243] S Vijayarani, Ms J Ilamathi, and Ms Nithya. Preprocessing techniques for text mining-an overview. *International Journal of Computer Science & Communication Networks*, 5(1):7–16, 2015.

[244] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, and Daan Wierstra. Matching networks for one shot learning. *Advances in neural information processing systems*, 29, 2016.

[245] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011.

[246] Congcong Wang and David Lillis. Multi-task transfer learning for finding actionable information from crisis-related messages on social media. *arXiv preprint arXiv:2102.13395*, 2021.

[247] Fei Wang, Mengqing Jiang, Chen Qian, Shuo Yang, Cheng Li, Honggang Zhang, Xiaogang Wang, and Xiaoou Tang. Residual attention network for image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164.

[248] Ruxin Wang, Tongliang Liu, and Dacheng Tao. Multiclass learning with partially corrupted labels. *IEEE transactions on neural networks and learning systems*, 29(6):2568–2580, 2017.

[249] Shuhui Wang, Yangyu Chen, Junbao Zhuo, Qingming Huang, and Qi Tian. Joint global and co-attentive representation learning for image-sentence retrieval. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 1398–1406.

[250] Tan Wang, Xing Xu, Yang Yang, Alan Hanjalic, Heng Tao Shen, and Jingkuan Song. Matching images and text with multi-modal tensor fusion and re-ranking. In *Proceedings of the 27th ACM international conference on multimedia*, pages 12–20.

[251] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 568–578.

[252] Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, and Subhojit Som. Image as a foreign language: Beit pretraining for all vision and vision-language tasks. *arXiv preprint arXiv:2208.10442*, 2022.

[253] Xiaolong Wang and Abhinav Gupta. Unsupervised learning of visual representations using videos. In *Proceedings of the IEEE international conference on computer vision*, pages 2794–2802.

[254] Jiaheng Wei, Zhaowei Zhu, Hao Cheng, Tongliang Liu, Gang Niu, and Yang Liu. Learning with noisy labels revisited: A study using real-world human annotations. *arXiv preprint arXiv:2110.12088*, 2021.

[255] Gary M Weiss. Mining with rarity: a unifying framework. *ACM Sigkdd Explorations Newsletter*, 6(1):7–19, 2004.

[256] Dirk Weissenborn, Oscar Täckström, and Jakob Uszkoreit. Scaling autoregressive video models. *arXiv preprint arXiv:1906.02634*, 2019.

[257] Michael J Widener and Wenwen Li. Using geolocated twitter data to monitor the prevalence of healthy and unhealthy food references across the us. *Applied Geography*, 54:189–197, 2014.

[258] Chuhan Wu, Fangzhao Wu, and Yongfeng Huang. Rethinking infonce: How many negative samples do you need? *arXiv preprint arXiv:2105.13003*, 2021.

[259] Jiancan Wu, Xiang Wang, Fuli Feng, Xiangnan He, Liang Chen, Jianxun Lian, and Xing Xie. Self-supervised graph learning for recommendation. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*, pages 726–735.

[260] Yaxiong Wu, Craig Macdonald, and Iadh Ounis. Partially observable reinforcement learning for dialog-based interactive recommendation. In *Proceedings of the 15th ACM Conference on Recommender Systems*, pages 241–251.

[261] Patrick Xia, Shijie Wu, and Benjamin Van Durme. Which* bert? a survey organizing contextualized encoders. *arXiv preprint arXiv:2010.00854*, 2020.

[262] Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. Simmim: A simple framework for masked image modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9653–9663.

[263] Yufei Xu, Qiming Zhang, Jing Zhang, and Dacheng Tao. Vitae: Vision transformer advanced by exploring intrinsic inductive bias. *Advances in neural information processing systems*, 34:28522–28535, 2021.

[264] I Zeki Yalniz, Hervé Jégou, Kan Chen, Manohar Paluri, and Dhruv Mahajan. Billion-scale semi-supervised learning for image classification. *arXiv preprint arXiv:1905.00546*, 2019.

[265] Chenxiao Yang, Qitian Wu, Jipeng Jin, Xiaofeng Gao, Junwei Pan, and Guihai Chen. Trading hard negatives and true negatives: A debiased contrastive collaborative filtering approach. *arXiv preprint arXiv:2204.11752*, 2022.

[266] Jheng-Hong Yang, Carlos Lassance, Rafael Sampaio De Rezende, Krishna Srinivasan, Miriam Redi, Stéphane Clinchant, and Jimmy Lin. Atomic: An image/text retrieval test collection to support multimedia content creation. In *Proceedings of the 46th International ACM SIGIR conference on research and development in information retrieval*, pages 2975–2984.

[267] Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alex Smola. Stacked attention networks for image question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 21–29.

[268] Yazhou Yao, Zeren Sun, Chuanyi Zhang, Fumin Shen, Qi Wu, Jian Zhang, and Zhenmin Tang. Jo-src: A contrastive approach for combating noisy labels. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5192–5201.

[269] Yu Yao, Tongliang Liu, Bo Han, Mingming Gong, Jiankang Deng, Gang Niu, and Masashi Sugiyama. Dual t: Reducing estimation error for transition matrix in label-noise learning. *Advances in neural information processing systems*, 33:7260–7271, 2020.

[270] Xinyue Ye, Shengwen Li, Xining Yang, and Chenglin Qin. Use of social media for the detection and analysis of infectious diseases in china. *ISPRS International Journal of Geo-Information*, 5(9):156, 2016.

[271] Zixuan Yi, Zijun Long, Iadh Ounis, Craig Macdonald, and Richard Mccreadie. Large multi-modal encoders for recommendation. In *arXiv preprint arXiv:2310.20343*, 2023.

[272] LIN Yong, Renjie Pi, Weizhong Zhang, Xiaobo Xia, Jiahui Gao, Xiao Zhou, Tongliang Liu, and Bo Han. A holistic view of label noise transition matrix in deep learning and beyond. In *The Eleventh International Conference on Learning Representations*.

[273] Atsuo Yoshitaka and Tadao Ichikawa. A survey on content-based retrieval for multimedia databases. *IEEE Transactions on Knowledge and Data Engineering*, 11(1):81–93, 1999.

[274] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014.

[275] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*, 2022.

[276] Chang Yue and Niraj K Jha. Ctrl: Clustering training losses for label error detection. *IEEE Transactions on Artificial Intelligence*, 2024.

[277] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6023–6032.

[278] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *International Conference on Machine Learning*, pages 12310–12320. PMLR.

[279] Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. Lit: Zero-shot transfer with locked-image text tuning. In

Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 18123–18133.

[280] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3):107–115, 2021.

[281] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3):107–115, 2021.

[282] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017.

[283] Lisai Zhang, Hongfa Wu, Qingcai Chen, Yimeng Deng, Zhonghua Li, Dejiang Kong, Zhao Cao, Joanna Siebert, and Yunpeng Han. Vldeformer: learning visual-semantic embeddings by vision-language transformer decomposing. *arXiv preprint arXiv:2110.11338*, 9, 2021.

[284] Tianyi Zhang, Felix Wu, Arzoo Katiyar, Kilian Q Weinberger, and Yoav Artzi. Revisiting few-sample bert fine-tuning. *arXiv preprint arXiv:2006.05987*, 2020.

[285] Xueting Zhang, Debin Meng, Henry Gouk, and Timothy M Hospedales. Shallow bayesian meta learning for real-world few-shot recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 651–660.

[286] Yin Zhang, Rong Jin, and Zhi-Hua Zhou. Understanding bag-of-words model: a statistical framework. *International journal of machine learning and cybernetics*, 1:43–52, 2010.

[287] Han Zhao, Xu Yang, Zhenru Wang, Erkun Yang, and Cheng Deng. Graph debiased contrastive learning with joint representation clustering. In *IJCAI*, pages 3434–3440.

[288] Zijia Zhao, Longteng Guo, Xingjian He, Shuai Shao, Zehuan Yuan, and Jing Liu. Mamo: Fine-grained vision-language representations learning with masked multimodal modeling. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1528–1538.

[289] Songzhu Zheng, Pengxiang Wu, Aman Goswami, Mayank Goswami, Dimitris Metaxas, and Chao Chen. Error-bounded correction of noisy labels. In *International Conference on Machine Learning*, pages 11447–11457. PMLR.

[290] Bolei Zhou, Agata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. Learning deep features for scene recognition using places database. *Advances in neural information processing systems*, 27, 2014.

[291] Bolei Zhou, Agata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. Learning deep features for scene recognition using places database. *Advances in neural information processing systems*, 27, 2014.

[292] Hong-Yu Zhou, Chixiang Lu, Sibei Yang, and Yizhou Yu. Convnets vs. transformers: Whose visual representations are more transferable? In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2230–2238.

[293] Haoran Zhu, Boyuan Chen, and Carter Yang. Understanding why vit trains badly on small datasets: An intuitive perspective. *arXiv preprint arXiv:2302.03751*, 2023.

[294] Qiu-Shi Zhu, Long Zhou, Jie Zhang, Shu-Jie Liu, Yu-Chen Hu, and Li-Rong Dai. Robust data2vec: Noise-robust speech representation learning for asr by combining regression and improved contrastive learning. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

[295] Zhiqiang Zou, Hongyu Gan, Qunying Huang, Tianhui Cai, and Kai Cao. Disaster image classification by fusing multimodal social media data. *ISPRS International Journal of Geo-Information*, 10(10):636, 2021.