THE PENNSYLVANIA STATE UNIVERSITY
SCHREYER HONORS COLLEGE

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

Exploring Hybrid Pre-Training and Fine-Tuning Strategies for Multimodal Transfer Learning in Cross-Modal Retrieval

JAYESH AGARWALA
SPRING 2024

A thesis
submitted in partial fulfillment
of the requirements
for a baccalaureate degree
in Computer Science
with honors in Computer Science

Reviewed and approved* by the following:

Wang-Chien Lee
Associate Professor of Computer Science and Engineering
Thesis Supervisor

Ting He
Associate Professor of Computer Science and Engineering
Honors Adviser

* Electronic approvals are on file.

ABSTRACT

Multimodal transfer learning offers a powerful solution for cross-modal retrieval tasks by leveraging knowledge across modalities. In this thesis, we explore a two-stage pre-training and fine-tuning approach within an existing multimodal transfer learning framework to improve model efficiency and adaptability. While we don't claim superiority in retrieval accuracy and robustness compared to traditional methods, our research provides valuable insights into optimizing performance for cross-modal retrieval tasks.

This exploration involves dividing the model into pre-training and fine-tuning stages. By investigating various configurations within this framework, we aim to identify strategies that can reduce training time and epochs, while also enhancing the model's ability to adapt to new data categories. Our experiments analyze the factors that influence performance in this two-stage approach, providing valuable guidance for future research in multimodal transfer learning.

This work contributes to advancing the design and optimization of cross-modal retrieval systems. By exploring segmentation strategies within existing models, our findings can inform the development of more efficient and adaptable retrieval systems for real-world applications.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I am sincerely grateful to my Thesis supervisor, Wang-Chien Lee, for his invaluable guidance and unwavering support throughout the course of my thesis. His expertise and mentorship have been instrumental in shaping my research journey and achieving significant milestones. I extend my heartfelt appreciation to Professor Meng-Fen Chiang for her insightful guidance and encouragement, which have helped me delve deeper into the intricacies of the topic and derive meaningful insights and results. I would like to express my sincere gratitude to my Honors Advisor, Ting He, for her invaluable assistance and steadfast support during my final year. Finally, I extend my deepest thanks to my family and friends for their unwavering support, encouragement, and understanding throughout this journey. Their love and encouragement have been a constant source of strength and inspiration.

Chapter 1

Introduction

1.1 Background and Motivation

Cross-modal retrieval (CMR) is a burgeoning field crucial for developing search systems that operate seamlessly across different modalities (e.g., text, images, video). CMR allows users to search for images using textual descriptions, retrieve videos matching an audio query, or discover text articles related to a photograph. This ability to bridge representational gaps between modalities opens up new possibilities for organizing, searching, and understanding our rich multimedia world. (Wang, Yin, Wang, Wu, & Wang, 2016)

Transfer learning plays a vital role in addressing the hurdles of CMR. Labeled training data for specific cross-modal tasks is often limited, hindering the performance of models, especially when new or unseen data categories emerge. Transfer learning allows models to leverage knowledge gleaned from extensive datasets, often from different domains, and apply that knowledge to new target tasks. This transfer of representations improves generalization abilities and allows models to adapt more rapidly to novel data categories.

Despite offering significant advantages, training complex CMR models can be a computationally intensive and time-consuming process. Additionally, ensuring that models smoothly adapt to new data categories remains a primary goal for real-world applications. Motivated by these challenges, our research focuses on strategies for improving the efficiency

and adaptability of CMR models. Specifically, we aim to reduce training time and computational demands, while simultaneously enhancing the capacity of models to quickly adjust to new data they encounter.

Improving the efficiency and adaptability of CMR models has far-reaching benefits across various domains. In e-commerce, accurate and responsive image search using text descriptions can streamline product discovery for customers. In healthcare, cross-modal retrieval systems that link medical images with related diagnostic reports can improve the efficiency and accuracy of clinical decision-making. The ability to learn faster and adapt to new knowledge is critical for developing intelligent multimodal applications that can navigate the complexities of real-world data.

1.2 Research Objectives

This thesis explores the potential of optimizing performance within the current state-of-the-art model for deep multimodal transfer learning (DMTL) in cross-modal retrieval (CMR) tasks. The model we reference is the one presented in the paper, "Deep Multimodal Transfer Learning for Cross-Modal Retrieval" (Zhen, Hu, Peng, Goh, & Zhou, 2022), which focus on transferring knowledge from labeled categories (source domain) to unlabeled categories (target domain) where the label sets are disjoint. Their method employs a joint learning strategy to assign pseudolabels to the target samples and leverages modality-specific networks to learn a shared semantic space, aiming to bridge the heterogeneity gap between modalities.

However, the original paper's primary emphasis lies on achieving high retrieval accuracy and robustness, potentially at the cost of training efficiency. In contrast, this thesis adopts a different perspective. We prioritize improving the efficiency and adaptability of CMR models while maintaining performance. This shift is motivated by the need for faster training times and the ability of models to adapt to new data categories often encountered in real-world applications.

Our core research objective is to explore a two-stage pre-training and fine-tuning approach within an existing multimodal transfer learning framework. By investigating various configurations within this framework, we aim to identify strategies that can:

- Reduce training time and computational resources required for training CMR models.
- Enhance the model's ability to adapt and learn from new categories of data encountered after initial training.

By achieving these objectives, we hope to contribute to the development of more efficient and adaptable CMR systems that can be readily deployed in real-world applications.

1.3 Contributions of the Thesis

This thesis investigated hybrid pre-training and fine-tuning strategies for effective knowledge transfer in cross-modal retrieval. Several key contributions emerge from the experimental results and analysis:

- **Importance of Joint Learning:** The findings demonstrate the advantages of training models on both source and target data simultaneously. This joint learning strategy facilitates efficient knowledge transfer and adaptation to unlabelled target domains.
- **Parameter Tuning Sensitivity:** Optimal model performance is highly dependent on the careful tuning of hyperparameters, specifically those governing the balance between source and target data emphasis. These parameters need to be adjusted in accordance with dataset characteristics.
- **Dataset-Driven Strategies:** The quantity and distribution of categories within the source dataset significantly influence knowledge transfer effectiveness. This highlights the need for dataset-aware strategies when selecting training data splits or designing augmentation techniques.

1.4 Scope and Organisation of the Thesis

This thesis investigates strategies for optimizing the efficiency and adaptability of multimodal transfer learning models within the context of cross-modal retrieval. We explore the potential of a two-stage pre-training and fine-tuning approach to reduce training time, enhance adaptability to new data categories, and ultimately contribute to the development of more practical cross-modal retrieval systems.

The thesis is organized as follows:

- **Background Work:** This section provides an overview of cross-modal retrieval, its significance, and the role of transfer learning in addressing CMR challenges. It also reviews relevant literature on existing optimization strategies in multimodal models.
- **Problem Formulation:** This section formally defines the problem addressed in the thesis, outlining objectives and the specific challenges of optimizing training efficiency and adaptability for multimodal transfer learning. It also discusses the scope of the research and any key assumptions.
- **Methods and Experimental Setting:** Here, we detail the proposed pre-training and fine-tuning strategies, along with any modifications to the existing model architecture. The experimental setup description includes datasets, evaluation metrics, and hyperparameter settings.
- **Experimental Results with Observations:** This section presents quantitative results of the experiments and a thorough analysis of performance relative to both traditional and state-of-the-art methods. Key observations and insights from the data are highlighted.
- **Discussion on Findings and Future Work:** We analyze the implications of the experimental results, discussing how the proposed approach addresses the initial problem. We identify limitations and outline potential avenues for further research and improvement.
- **Conclusion:** The thesis concludes with a summary of the research contributions, a restatement of the main findings, and their significance within the broader context of cross-modal retrieval.

Chapter 2

Theoretical Foundations and Related Work

2.1 Background

This section provides a comprehensive overview of the background concepts and techniques relevant to the thesis, drawing upon the insights presented in the reference paper "Deep Multimodal Transfer Learning for Cross-Modal Retrieval".

2.1.1 Cross-Modal Retrieval (CMR)

CMR is an emerging field in information retrieval that deals with searching for relevant information across different modalities. In simpler terms, it allows users to retrieve information from one data type (modality) based on a query from another modality. Classic examples include searching for images using textual descriptions or vice versa, or finding videos that match an audio query.

The ability to bridge the gap between different modalities presents significant opportunities for developing more intuitive and flexible search systems across various domains.

2.1.2 Transfer Learning

Transfer learning is a machine learning technique that leverages knowledge gained from a source task to improve performance on a related target task. This approach is particularly beneficial when labeled data for the target task is scarce or expensive to acquire. In the context of CMR, transfer learning can be employed to transfer knowledge from labeled data in one domain (source domain) to a different domain (target domain) where labels are unavailable. This enables models to learn transferable representations that can be adapted to unseen categories in the target domain, improving overall retrieval accuracy.

There are various categories of transfer learning approaches, but this thesis focuses on the application of transductive transfer learning for CMR tasks. In transductive transfer learning, both the source and target tasks are the same (i.e., CMR), but the data distributions differ between the source and target domains due to the lack of labels in the target domain.

2.1.3 Labeled and Unlabeled Data Categories

The effectiveness of CMR models heavily relies on the quality and quantity of training data. Training data typically consists of labeled and unlabeled categories. Labeled data refers to samples where each data point is associated with a corresponding category label. This label information provides crucial supervision for the model to learn the underlying relationships between modalities. Unlabeled data, on the other hand, lacks explicit category labels. While unlabeled data can still be informative for learning general patterns within a modality, it presents challenges for tasks like CMR that require identifying semantic relationships across modalities.

The original paper addresses the challenge of knowledge transfer in CMR scenarios where the source and target domains have disjoint label sets (Zhen et al., 2022). This means that the categories present in the labeled source domain are entirely different from the categories in the unlabeled target domain. This distinction is crucial because traditional transfer learning methods often assume that the source and target tasks share some common categories.

2.1.4 Parameter Tuning

The performance of machine learning models, including those for CMR, is highly sensitive to the hyperparameters chosen during training. Hyperparameters are essentially the settings that control the learning process of the model, but they are not directly learned from the data. Common examples of hyperparameters in deep learning models include learning rates, optimizer configurations, and network architectures.

Finding the optimal hyperparameter configuration is crucial for achieving good performance. Grid search involves systematically evaluating a range of possible values for each hyperparameter, while random search randomly samples hyperparameter values from a predefined search space.

2.1.5 Different Tasks in CMR

CMR encompasses a wide range of tasks depending on the specific modalities involved. Here are the tasks we will be analyzing in this thesis:

- Image-to-Text Retrieval: Given a textual description, the task is to retrieve relevant images from a collection.
- Text-to-Image Retrieval: Given an image, the task is to retrieve textual descriptions that accurately represent the image content.

2.1.6 Zero-Shot Cross-Modal Retrieval

Zero-shot learning (ZSL) is a challenging paradigm where a model must recognize or classify instances from categories that were unseen during training. In cross-modal retrieval, a zero-shot setting implies that the model is trained exclusively on a set of source categories and then directly tested on a disjoint set of target categories. This presents a significant challenge as the model has no prior exposure to any examples from the target categories.

This thesis investigates cross-modal retrieval in a zero-shot setting. Specifically, the following approach is employed:

- Training: The model is trained on a labelled source dataset containing image-text pairs from a specific set of categories (e.g., "cat," "dog," "flower").
- Testing: The model's ability to generalize to unseen target categories is evaluated. The model is tested on image-text pairs belonging to categories that were not present in the training data (e.g., "airplane," "car," "building").

This zero-shot setting mirrors real-world scenarios where a model trained on a limited set of data must adapt to new, potentially unrelated categories. To succeed in this context, effective

knowledge transfer strategies are crucial for maximizing the model's ability to leverage information learned from the source domain and apply it to the unseen target domain.

2.2 Related Work

The paper "Deep Multimodal Transfer Learning for Cross-Modal Retrieval" by Liangli Zhen, Peng Hu, Xi Peng, Rick Siow Mong Goh, and Joey Tianyi Zhou addresses the challenge of knowledge transfer across different modalities in scenarios where the source and target domains have disjoint label sets. Their proposed deep multimodal transfer learning (DMTL) approach utilizes a joint learning strategy with modality-specific neural networks to create a shared semantic space for different modalities. This framework demonstrates significant improvements in retrieval accuracy for cross-modal tasks, advancing the state of the art.

The work presented in the paper provides a strong foundation for our research. Their exploration of transfer learning with disjoint label sets motivates our investigation of pre-training and fine-tuning strategies to optimize model efficiency and adaptability within this framework.

2.2.1 DMTL Framework Overview

The proposed Deep Multimodal Transfer Learning (DMTL) framework addresses the challenge of transferring knowledge from labeled categories in a source domain to unlabeled categories in a target domain for cross-modal retrieval (CMR). This scenario is particularly

interesting because the label sets of the source and target domains are disjoint, meaning they do not share any common categories.

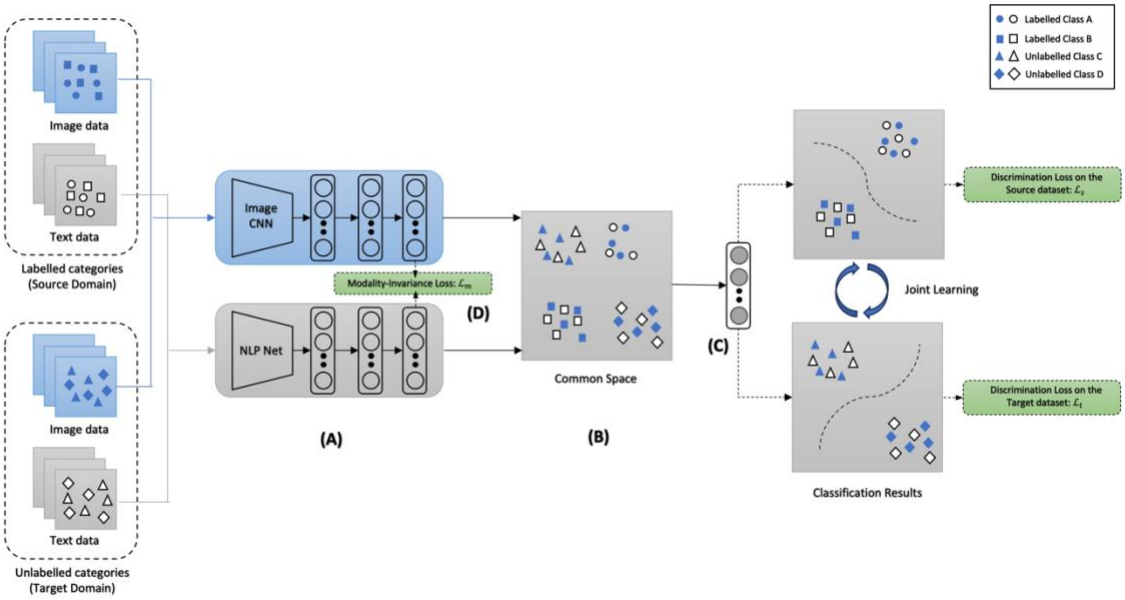

Figure 1 General Framework for DMTL method (Zhen et al., 2022) © 2022 IEEE

Here's a breakdown of the key components in the DMTL framework:

- **Modality-Specific Networks:** Framework employs two separate convolutional neural networks (CNNs) for image data and a natural language processing network (NLP Net) for text data as represented in Figure 1(A). These networks act as encoders, aiming to learn modality-specific representations for the input data.

- **Shared Semantic Space:** The CNN and NLP Net representations are projected into a shared latent space using fully connected layers as represented in Figure 1(B). This space aims to capture the underlying semantic relationships between different modalities.
- **Joint Learning with Pseudolabels:** To bridge the gap between labeled and unlabeled data, DMTL incorporates a joint learning strategy. For each unlabeled sample in the target domain, a pseudolabel is assigned and iteratively refined during the training process. This pseudolabel serves as a temporary supervisory signal, guiding the network to learn transferable features even for unseen categories.
- **Category Information Exploitation:** The framework leverages category information from the labeled source domain to guide the learning of pseudolabels for the unlabelled target domain. This is achieved by feeding the shared representations from both domains into a linear classifier for label prediction, even though the target categories are unknown as represented in Figure 1(C).
- **Heterogeneity Gap Reduction:** By enforcing specific properties in the shared space, DMTL aims to reduce the heterogeneity gap between different modalities (Figure 1(D)). This is achieved by minimizing the distance between samples from the same category (homoinstances) and maximizing the distance between samples from different categories (heteroinstances).

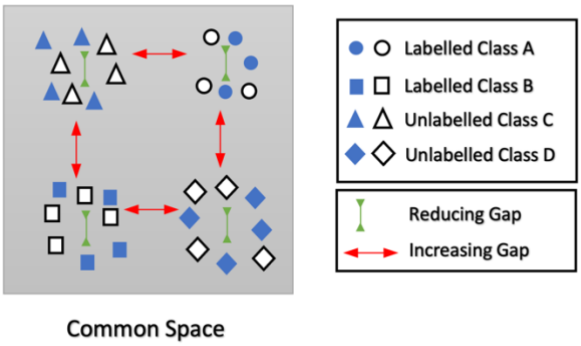

Figure 2 Hetrogenity Gap Reduction in Cross-Modal Retrieval

Overall, the DMTL framework combines modality-specific encoders, a shared semantic space, joint learning with pseudolabels, category information exploitation, and heterogeneity gap reduction to achieve knowledge transfer for CMR in the target domain with disjoint categories.

2.2.2 Objective Function

All the equations mentioned in this section is adapted from the formulation presented in Zhen et al. (2021) and combines several key strategies for effective cross-modal retrieval in scenarios with unlabelled target data.

The objective function in DMTL is:

$$\mathcal{L} = \mathcal{L}_m + \lambda_1 \mathcal{L}_s + \lambda_2 \mathcal{L}_t \quad (1)$$

- $\boldsymbol{\mathcal{L}_m}$**:** Modality-invariance loss. Aims to bridge the gap between different modalities (e.g., image and text) by ensuring that semantically related samples across modalities are brought closer in a shared embedding space.
- $\boldsymbol{\mathcal{L}_s}$**:** Discrimination loss for the labeled source data. Encourages the model to learn discriminative features by leveraging the label information in the source domain.
- $\boldsymbol{\mathcal{L}_t}$**:** Discrimination loss for the unlabeled target data. Seeks to learn discriminative features for the target domain by using pseudolabels and minimizing the difference between these pseudolabels over successive iterations.
- $\boldsymbol{\lambda_1, \lambda_2}$**:** Tradeoff parameters, controlling the relative importance of each loss term.

Discrimination loss for labelled source data is defined as:

$$\mathcal{L}_s = \frac{1}{m}\sum_{i=1}^{m}(||Ph^{\alpha}(s_i^{\alpha}) - y_i||_2 \; + \; Ph^{\beta}(s_i^{\beta}) - y_i||_2) \tag{2}$$

- $\boldsymbol{P}$**:** Weight matrix of a linear classifier used in the discriminative loss terms.
- $\boldsymbol{h^{\alpha}, h^{\beta}}$**:** Transformation functions for image and text samples. These play a central role in mapping image and text samples into the shared embedding space.
- $\boldsymbol{s^{\alpha}, s^{\beta}}$**:** Image and text samples, respectively.
- $\boldsymbol{y}$**:** Semantic label vectors of samples
- $\boldsymbol{m}$**:** Instances of image-text pair in source domain

Discrimination loss for unlabeled target data is defined as:

$$\mathcal{L}_t = \frac{1}{n}\sum_{j=1}^{n}(||Ph^{\alpha}(x_j^{\alpha}) - z_j^{\alpha}||_2 \; + \; Ph^{\beta}(x_i^{\beta}) - z_j^{\beta}||_2) \tag{3}$$

- $\boldsymbol{x^{\alpha}, x^{\beta}}$**:** Image and text samples, respectively.
- $\boldsymbol{z^{\alpha}, z^{\beta}}$**:** Pseudolabels for the unlabelled target data, representing the degree of similarity/dissimilarity between the target data and the categories in the source domain.
- $\boldsymbol{n}$**:** Instances of image-text pair in source domain

During training, the pseudolabels of the samples in the target data set will be updated at each iteration by:

$$z_j^{\alpha} = Ph^{\alpha}\left(x_j^{\alpha}\right) \quad (4)$$

$$z_j^{\beta} = Ph^{\beta}\left(x_j^{\beta}\right) \quad (5)$$

2.2.3 Optimization Goals:

To achieve effective cross-modal retrieval with knowledge transfer to an unlabelled target domain, the optimization process outlined in Section 2.2.2 pursues several interconnected goals:

- **Cross-Modal Matching:** Minimize the gap between different modalities (image and text) through the modality-invariance loss ($\mathcal{L}_m$). This encourages the model to find representations where related image and text samples are close together.
- **Knowledge Transfer:** Utilize labeled source data ($\mathcal{L}_s$) to guide model learning and transfer knowledge to the target domain, while also incorporating information from unlabeled target data ($\mathcal{L}_t$).
- **Discriminative Feature Learning:** Learn features that effectively distinguish between different categories, particularly focusing on the unlabeled target data and the knowledge transferred from the source domain.

Chapter 3

Problem Formulation

Cross-Modal Retrieval (CMR) tackles the fundamental problem of enabling seamless search across different modalities like images and text. The central challenge stems from the fact that these modalities exist in distinct representational spaces, making direct comparisons difficult. To bridge this gap, CMR aims to learn transformations (h^{α} for images, h^{β} for text) that map samples from both modalities into a shared embedding space. In this space, the goal is for semantically related items to cluster together, regardless of whether they are images or text.

A major hurdle in real-world CMR applications is the limited availability of labeled data for new or niche domains. This thesis addresses this challenge by proposing a transfer learning approach specifically designed to leverage knowledge from a labeled source domain and effectively transfer it to a target domain with unlabeled data. This strategy is crucial for scenarios where the categories of interest in the target domain might be entirely different from those with readily available labels.

Following the notation of the "Deep Multimodal Transfer Learning for Cross-Modal Retrieval" Paper, we will formalize the problem setup:

- Source Domain
    - Contains $\boldsymbol{C_s}$ labelled categories
    - $\boldsymbol{m}$ instances of image-text pair denoted as $\boldsymbol{S} = \{(\boldsymbol{s_i^{\alpha}}, \boldsymbol{s_i^{\beta}})\}_{\boldsymbol{i=1}}^{\boldsymbol{m}}$

- Each pair $(s_i^{\alpha}, s_i^{\beta})$ possesses a label vector $\boldsymbol{y_i} = [\boldsymbol{y_{1i}}, \boldsymbol{y_{2i}}, \ldots, \boldsymbol{y_{C_s i}}] \in \mathbb{R}^{\boldsymbol{C_s}}$, indicating which category the sample belongs to. If the i[th] instance belongs to the k[th] category, $\boldsymbol{y_{ki}} = \boldsymbol{1}$, otherwise $\boldsymbol{y_{ki}} = \boldsymbol{0}$

- **Target Domain**
  - Contains $\boldsymbol{C_x}$ unlabeled new categories
  - $\boldsymbol{n}$ instances of image-text pair denoted as $\boldsymbol{X} = \{(\boldsymbol{x_j^{\alpha}}, \boldsymbol{x_j^{\beta}})\}_{\boldsymbol{j=1}}^{\boldsymbol{n}}$
  - Samples in target domain are without labels

Importantly, target domain categories have no overlap with the source domain categories, signifying a need for knowledge transfer to unseen categories.

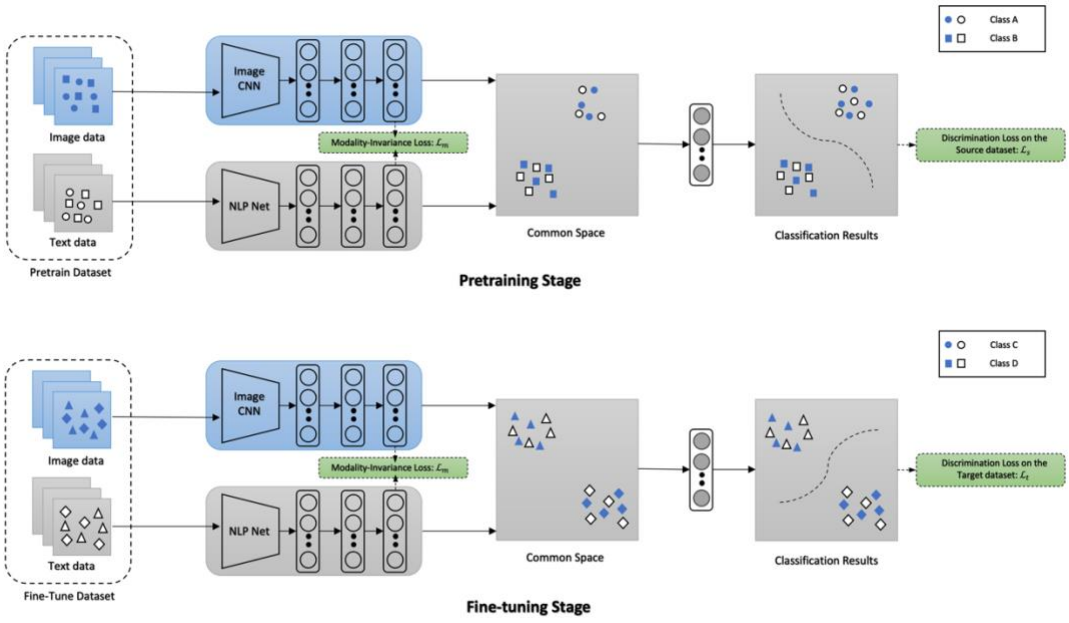

Figure 3 Two-Stage training approach for Cross-Modal Retrieval

To tackle this, the thesis adopts a two-stage training approach as illustrated in Figure 3. The first stage, pre-training, involves training a cross-modal model on a large and potentially

more generic multimodal dataset. This stage is essential for the model to learn robust cross-modal correspondences that generalize well. Even if the pre-training data is not a perfect domain match for the target data, it provides a strong foundation for mapping into a shared representation space. The second stage, fine-tuning, focuses on tailoring the pre-trained model to the specifics of the target dataset, including its unique data distribution and categories.

This two-stage approach offers several key advantages. Pre-training on a diverse dataset significantly improves model performance and generalization, mitigating overfitting, which is a critical concern when the target dataset is small. Additionally, starting from a pre-trained state leads to significantly faster convergence during fine-tuning, saving training time and computational resources. Finally, this method is resource-efficient. In cases where the target dataset is relatively small, it leverages the knowledge learned during pre-training, compensating for the limited data in the target domain.

Chapter 4

Methods and Experimental Settings

4.1 Methods - Experimental Progression for Enhanced Cross-Modal Retrieval

This thesis investigates a deep multimodal transfer learning approach to facilitate cross-modal retrieval (CMR), specifically focusing on the challenge of transferring knowledge from a labeled source domain to a target domain with unlabeled data and disjoint categories. An iterative experimental design was adopted to explore various strategies and analyze their impact on CMR performance in this setting.

4.1.1 Version 1: Pre-training & Zero-Shot Testing

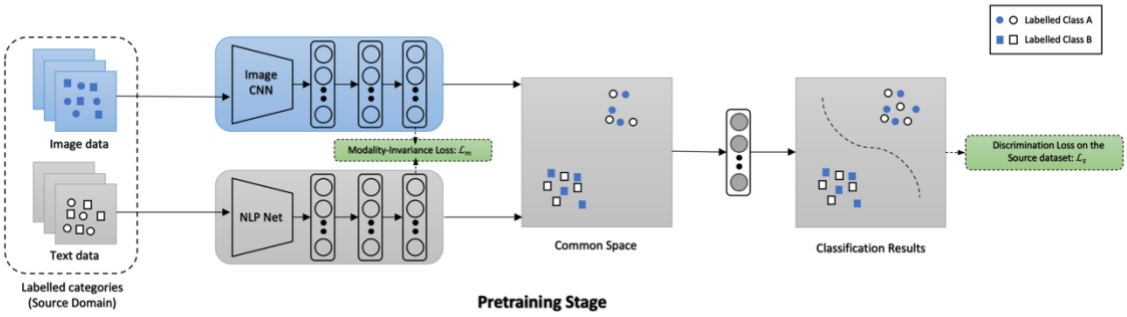

Figure 4. Version 1 - Pretraining stage trained only on source labelled dataset

- **Procedure:** The initial model was pre-trained solely on the source labelled dataset using the objective function $\mathcal{L} = \mathcal{L}_m + \lambda_1 \mathcal{L}_s$. It was then directly evaluated on the target domain without any exposure to target data.
- **Reasoning:** This setup served as a baseline to assess whether the model acquires any generalizable cross-modal knowledge during pre-training even without explicit knowledge transfer techniques. A degree of performance in this zero-shot scenario would indicate the potential for transfer learning.
- **Aim:** Establish the value of pre-training and gauge if source domain knowledge offers at least a basic starting point for target domain performance.

4.1.2 Version 2: Pre-training & Fine-tuning (with Labelled Target Data)

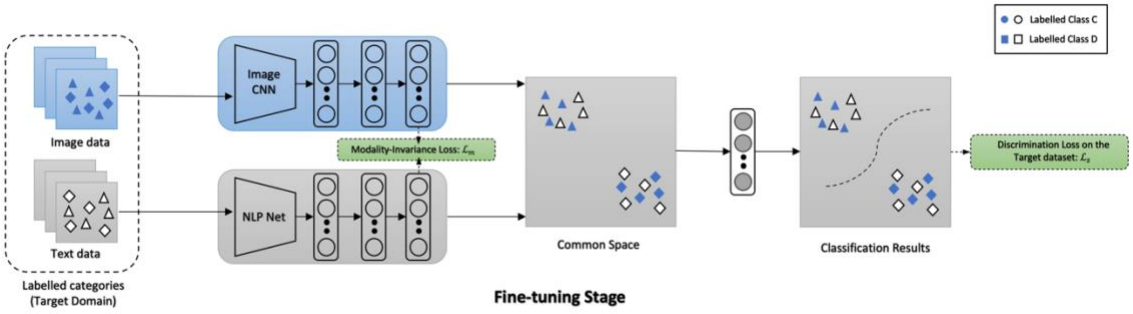

Figure 5. Version 2 - Fine-tuning stage trained only on unlabelled target dataset

- **Procedure:** The pre-trained model (from Version 1) underwent fine-tuning with the addition of labeled target data, using the same objective function $\mathcal{L} = \mathcal{L}_m + \lambda_1 \mathcal{L}_t$. Here the formulation of $\mathcal{L}_t$ is same as $\mathcal{L}_s$ because we know the labels of target data.

- **Reasoning:** While not directly aligned with the unlabelled target data problem, this version is crucial for verifying model correctness. The model performance is expected to be higher and not to be compared with other versions. Success here suggests that the model has the capacity to adapt to new data when labels are available, strengthening the validity of subsequent variations.
- **Aim:** Validate the model architecture and optimization pipeline, ensuring the ability to learn within the transfer learning framework.

4.1.3 Version 3: Pre-training, Fine-tuning (with Pseudolabels)

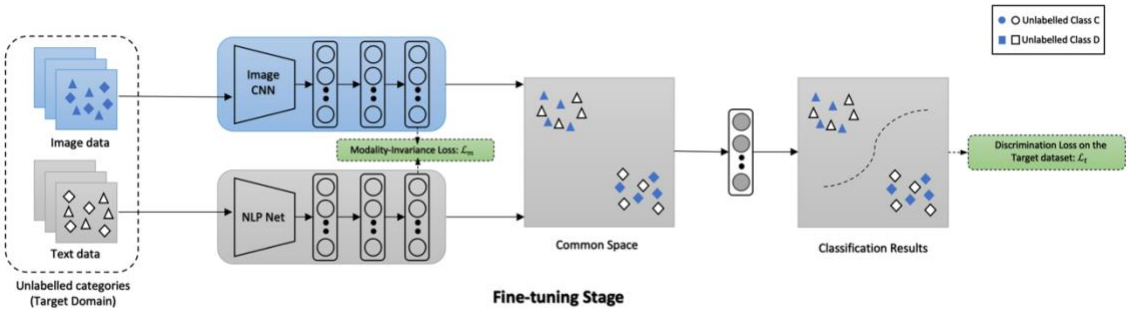

Figure 6. Version 3 - Fine-tuning stage trained on unlabelled target dataset

- **Procedure:** Pre-training remained as in Version 1. Fine-tuning then incorporated the unlabelled target data with pseudolabels, modifying the objective function to $\mathcal{L} = \mathcal{L}_m + \lambda_1 \mathcal{L}_s + \lambda_2 \mathcal{L}_t$.

- **Reasoning:** This introduces the core idea of knowledge transfer using pseudolabels to represent target domain semantics based on source domain knowledge. It aligns with the thesis's focus on scenarios where target labels are unavailable.
- **Aim:** Investigate the effectiveness of pseudolabels as a substitute for true labels and quantify gains in target domain performance due to this transfer strategy.

4.1.4 Version 4: Full Source Data for Pre-training & Fine-tuning

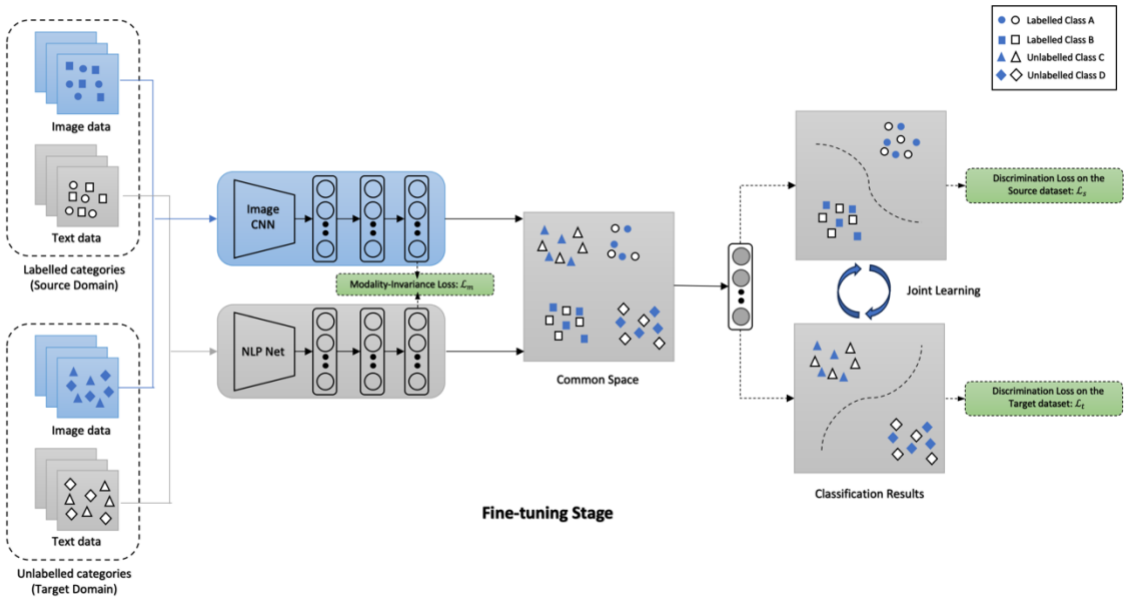

Figure 7. Version 4, Version 5, Version 6 - Fine-tuning stage trained on labelled source dataset and unlabelled target dataset utilizing joint-learning strategy

- **Procedure:** Pre-training used the entire source dataset. Fine-tuning continued to leverage pseudolabels on the unlabelled target data while also incorporating the entire source dataset, maintaining $\mathcal{L} = \mathcal{L}_m + \lambda_1 \mathcal{L}_s + \lambda_2 \mathcal{L}_t$.
- **Reasoning:** Exposing the model to more source data during both stages offers potential improvement in the quality of representation learning and pseudolabels.
- **Aim:** Determine if maximizing source domain knowledge benefits cross-modal alignment and overall performance in the transfer setting.

4.1.5 Version 5: Category-Split Source Data

- **Procedure:** Source data was split by category for pre-training and fine-tuning. Fine-tuning still used pseudolabels with the target data. Objective functions remained as in Version 4: $\mathcal{L} = \mathcal{L}_m + \lambda_1 \mathcal{L}_s + \lambda_2 \mathcal{L}_t$.
- **Reasoning:** This probes the model's ability to transfer knowledge even when pre-training and fine-tuning source categories are non-overlapping. This simulates a more realistic and challenging scenario where the pre-training data might not directly cover the categories of interest.
- **Aim:** Examine the robustness of knowledge transfer when the source and target categories are less aligned.

4.1.6 Version 6: Randomly-Split Source Data

- **Procedure:** Pre-training and fine-tuning utilized randomly divided halves of the source dataset. The rest of the setup mirrored Version 4: $\mathcal{L} = \mathcal{L}_m + \lambda_1 \mathcal{L}_s + \lambda_2 \mathcal{L}_t$.
- **Reasoning:** This acts as a control for Version 5, demonstrating whether gains are due to category-specific transfer or simply from increased source data.
- **Aim:** Isolate the impact of category alignment on knowledge transfer vs. the effect of having more source data during both training stages.

4.2 Experimental Setting

4.2.1 Dataset – Wikipedia Dataset

The primary dataset for this research is derived from Wikipedia's "featured articles" (Pereira et al, 2014). It offers a rich collection of 2866 image-text pairs, each encompassing a single image and a corresponding text description (several paragraphs) related to the image content. The dataset covers ten high-level semantic categories, such as art, history, and sports, providing a diverse range of content for model training and testing. Following established practice, the dataset is divided into a training set of 2173 pairs and a testing set of 693 pairs to ensure consistency and facilitate comparison with original paper.

4.2.2 Experimental Setup

The experimental design centers on transductive transfer learning, a scenario where the model has access to unlabelled target data during training but not the true category labels. This setting simulates real-world challenges where knowledge transfer is required to unseen domains with potentially new or unlabelled categories.

To evaluate model performance under these conditions, both the training and testing sets from Wikipedia dataset are randomly split into source and target subsets. Each subset encompasses 50% of the original categories within that dataset. This approach ensures that models must generalize their knowledge and transfer it effectively to unseen categories during the testing phase.

For feature representation, consistency with the original paper is maintained by employing VGGNet (Simonyan, & Zisserman, 2014) for image features and Doc2Vec (Lau, & Baldwin, 2014) for text features. VGGNet, a widely used convolutional neural network architecture, has proven effective in learning robust image representations. Doc2Vec, a word embedding model, captures the semantic relationships between words within the text descriptions.

On top of these feature extractors, a multi-layer perceptron (MLP) architecture is implemented. The MLP consists of three fully-connected layers with ReLU (Rectified Linear Unit) (Nair, & Hinton, 2010) activation functions. The number of hidden units in these layers is 4096, 4096, and 512, respectively (Zhen et al., 2022). These layers project the high-dimensional feature vectors obtained from VGGNet and Doc2Vec into a shared embedding space, allowing for effective cross-modal retrieval.

The entire model is trained end-to-end using the PyTorch framework. The Adam optimizer (Kan, Shan, Zhang, Lao, & Chen, 2014), a popular optimization algorithm for deep learning models, is employed with a learning rate of 10^{-4}. A batch size of 100 samples is used for training, and the maximum number of training epochs is set to 50. These hyperparameters were chosen based on a combination of empirical evaluation and reference to established practices in deep learning research.

4.2.3 Evaluation Metrics

To assess the effectiveness of the proposed methods in CMR tasks, primary evaluation metric used is mean average precision (mAP) (Rasiwasia et al., 2010).

mAP is a widely adopted and robust metric in CMR tasks. It considers both the precision of retrieved results and their ranking order. The calculation of mAP involves iterating over all queries in the test set. For each query, the model retrieves a set of nearest neighbors from the target data based on their similarity in the shared embedding space. Precision is calculated at various ranking thresholds (e.g., top 1, top 5, top 10 retrieved items). Essentially, precision measures the proportion of relevant items within the retrieved set at a specific ranking position. Mean Average Precision (mAP) is then obtained by averaging the precision values across all ranking thresholds and all queries in the test set.

Here's the mathematical equation for mAP:

$$AP = \frac{1}{R}\sum_{r=1}^{N} P(r)\sigma(r)$$

where:

- **R:** Number of relevant items for a specific query
- **N:** Total number of samples in the target data
- $\sigma(\boldsymbol{r})$**:** Indicator function (1 if the r^{th} retrieved item is relevant to the query, 0 otherwise)
- $\boldsymbol{P}(\boldsymbol{r})$: Precision at ranking position r

Chapter 5

Experiment Results and Observations

The experimental results presented in this section provide crucial evidence for evaluating the effectiveness of the proposed cross-modal retrieval methods. By analyzing performance across model versions (summarized in Table 1), key trends are identified that illuminate the impact of specific techniques and parameter choices. These observations inform understanding of the strengths and potential limitations of the proposed approach within the context of knowledge transfer for unlabelled target domains.

5.1 Analysis of Model Versions

5.1.1 Version 1: Pretraining on Source Labelled Dataset Only

- **Summary:** This version establishes a baseline performance (mAP: 0.300 ± 0.038) without any knowledge transfer to the target domain. It serves as a reference point to gauge the effectiveness of subsequent versions that incorporate transfer learning strategies.
- **Observations:** The mAP score is considerably lower compared to the original paper's model (mAP: 0.523 ± 0.066). This indicates that directly applying the model trained on the source domain to the target domain (zero-shot learning) yields limited performance.

- **Insights:** This finding aligns with the well-known challenges of directly applying models trained on a source domain to a target domain with different data distributions. The model struggles to generalize effectively without some adaptation to the target data.

5.1.2 Version 2: Fine-Tuning on Target Labelled Dataset

- **Summary:** This version introduces fine-tuning on the target labelled dataset after pretraining on the source labelled dataset. It achieves a significant improvement in mAP (0.815 ± 0.030) compared to Version 1, demonstrating the benefit of target-specific adaptation.
- **Observations:** The mAP score surpasses the original paper's model, suggesting that fine-tuning on the target labelled data can be highly effective when labelled target data is available. This highlights the importance of target-specific adjustments for performance gains.
- **Insights:** This confirms the effectiveness of transfer learning when the target domain has labelled data. The model leverages the knowledge learned from the source domain and refines it on the target domain, leading to superior performance. The model performance was expected to be higher and not to be compared with other versions because it doesn’t align with our original problem setting of transferring knowledge to unlabelled target dataset.

Table 1 Performance comparison in terms of the (Mean ± Std) mAP scores on Wikipedia dataset over ten times of Monte Carlo simulations

Version	Stage	Parameter	Image => Text	Text => Image	Average	Avg. Epoch
Original	Source (Zero Shot)	$\lambda_1 = 0.8$	0.305 ± 0.042	0.295 ± 0.034	0.300 ± 0.038	18
	Source + Target	$\lambda_1 = 0.5$ $\lambda_2 = 4.0$	**0.502 ± 0.067**	**0.545 ± 0.070**	**0.523 ± 0.066**	**20**
1	Pretraining	$\lambda_1 = 0.8$	0.305 ± 0.042	0.295 ± 0.034	0.300 ± 0.038	18
2	Pretraining	$\lambda_1 = 0.8$	0.305 ± 0.042	0.295 ± 0.034	0.300 ± 0.038	18
	Fine-Tuning	$\lambda_1 = 3.0$	0.711 ± 0.040	0.920 ± 0.025	0.815 ± 0.030	13.4
3	Pretraining	$\lambda_1 = 0.8$	0.305 ± 0.042	0.295 ± 0.034	0.300 ± 0.038	18
	Fine-Tuning	$\lambda_2 = 2.0$	0.468 ± 0.073	0.467 ± 0.072	0.468 ± 0.072	19.9
4	Pretraining	$\lambda_1 = 0.8$	0.305 ± 0.042	0.295 ± 0.034	0.300 ± 0.038	18
	Fine-Tuning	$\lambda_1 = 1.0$ $\lambda_2 = 2.5$	**0.479 ± 0.064**	**0.514 ± 0.067**	**0.497 ± 0.065**	**13.4**
5	Pretraining	$\lambda_1 = 0.5$	0.269 ± 0.033	0.271 ± 0.021	0.270 ± 0.026	12.5
	Fine-Tuning	$\lambda_1 = 0.8$ $\lambda_2 = 3.2$	0.476 ± 0.075	0.504 ± 0.079	0.490 ± 0.076	15.3
6	Pretraining	$\lambda_1 = 1.2$	0.291 ± 0.028	0.284 ± 0.028	0.287 ± 0.028	28.2
	Fine-Tuning	$\lambda_1 = 0.8$ $\lambda_2 = 2.5$	0.470 ± 0.064	0.496 ± 0.069	0.483 ± 0.066	17.8

5.1.3 Version 3: Fine-Tuning on Target Unlabelled Dataset with Pseudolabels

- **Summary:** This version explores using pseudolabels for fine-tuning on the target unlabelled dataset. However, the mAP score (0.468 ± 0.072) falls below both Version 1 and the original paper's model.
- **Observations:** The performance is lower than expected, indicating that using pseudolabels for fine-tuning on the target unlabelled data not be as effective as using labelled target data (as in Version 2).
- **Insights:** There are two possible explanations. First, the quality of pseudolabels generated from the source model might be insufficient for effective fine-tuning on the target domain. Second, jointly training with source and target data, as in the original paper's model, might be crucial for successful knowledge transfer, even when using labelled data for the target domain (as opposed to pseudolabels).

5.1.4 Version 4: Fine-Tuning on Full Source Labelled + Target Unlabelled Dataset

- **Summary:** This version showcases the best performance among models (excluding the original paper's) with an mAP of 0.497 ± 0.065. It achieves a good balance between training speed and adaptability, fulfilling thesis goals.
- **Observations:** While it performs slightly lower than the original paper's model, this version offers a significant advantage in training efficiency by leveraging both source and

target data during fine-tuning. This aligns with thesis objectives of achieving fast training and adaptability.

- **Insights:** The inclusion of the target unlabelled data during fine-tuning likely helps the model adapt to the target domain to some extent, even without explicit labels. This suggests that the model can learn from the structure and distribution of the target data, even if the labels are unavailable.

5.1.5 Version 5: Fine-Tuning with Half Source Categories + Target Unlabelled Dataset

- **Summary:** This version investigates the impact of using a limited set of source categories during fine-tuning. The mAP score (0.490 ± 0.076) is slightly lower compared to Version 4.
- **Observations:** The performance dip suggests that using a reduced set of source categories for fine-tuning might hinder the model's ability to transfer knowledge effectively to the target domain.
- **Insights:** Fewer source categories might limit the variety of features the model can learn from and transfer to the target domain. This underlines the importance of having a sufficient number of source categories during fine-tuning for optimal performance.

5.1.6 Version 6: Fine-Tuning with Half Source Labelled + Target Unlabelled Dataset

- **Summary:** This version explores splitting the source data in half for pretraining and fine-tuning, along with target unlabelled data. The mAP score (0.483 ± 0.066) is the lowest among Versions 4-6.
- **Observations:** The performance decline and increase in training epochs suggest challenges with this approach. This may be due to the random split potentially creating suboptimal data distributions for either pretraining or fine-tuning stages. Interestingly, the pretraining result was better than Version 5. This suggests the split might have provided more diverse categories during pretraining, but the limited data for fine-tuning could have hampered performance.
- **Insights:** Source data quantity and diversity seem crucial for both pretraining and fine-tuning. Random splitting might create imbalances. This emphasizes the need for a sufficient amount of source data, even when adapting to unlabelled target data. Splitting the source data based on category similarity could potentially improve results, ensuring a well-distributed variety of features for both training stages.

5.2 Parameter Tuning Trends

This section analyzes the impact of parameter tuning on model performance, focusing specifically on the weighting between λ_1 ($\mathcal{L}_s$ parameter) and λ_2 ($\mathcal{L}_t$ parameter). By examining trends in heatmaps and mAP scores across different model versions, we can gain insights into

how these parameters influence knowledge transfer and adaptation to the unlabelled target domain.

5.2.1 Original Version

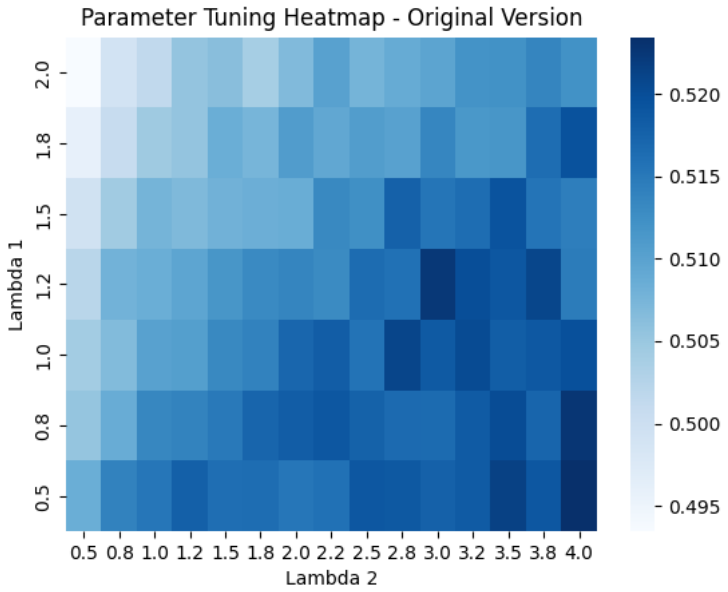

Figure 8 Heatmap of Original Version: Impact of λ_1 and λ_2 on mAP scores

The heatmap for the original model reveals a positive correlation between accuracy and higher values of λ_2 ($\mathcal{L}_t\ parameter$). Conversely, increasing λ_1 ($\mathcal{L}_s\ parameter$) appears to have a lesser or slightly negative impact on performance. This underscores the importance of prioritizing the model's ability to learn effectively from pseudolabels on the target dataset. By giving more weight to the unseen loss, the joint-learning process successfully guides knowledge transfer to the target domain, subsequently improving the overall mAP score. Aligning with this finding, we selected the λ_1 as 0.5 and λ_2 as 4.0 in the original model.

5.2.2 Version 4 (Full Source Pretraining; Full Source + Target Fine-tuning)

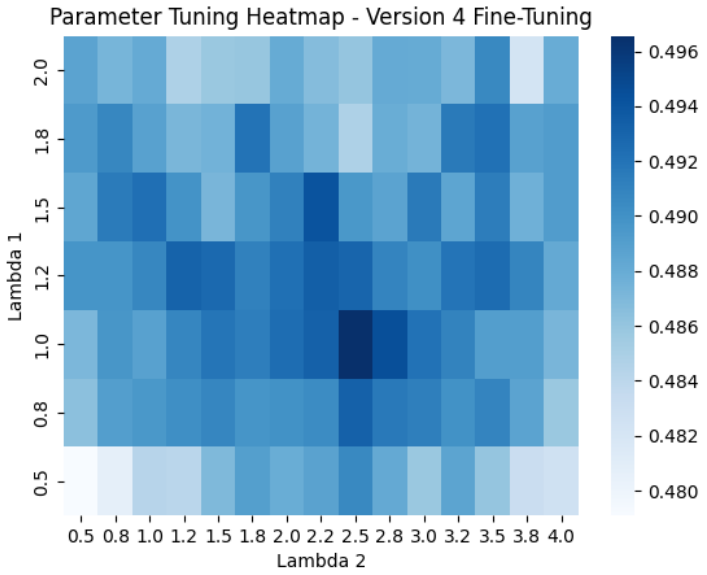

Figure 9 Heatmap of Version 4: Impact of λ_1 and λ_2 on mAP

In Version 4, while a higher λ_2 still benefits performance, the heatmap suggests the need for balance between both parameters. Excessively increasing λ_2 while heavily suppressing λ_1 potentially harms the model's ability to consolidate the knowledge acquired during pretraining. This implies that while knowledge transfer to the target domain is crucial, it's also important to retain and leverage the information learned from the labelled source dataset. Aligning with this finding, we selected the λ_1 as 1.0 and λ_2 as 2.5 in the version 4 model.

5.2.3 Version 5 (Category-Split Source Data + Target Data)

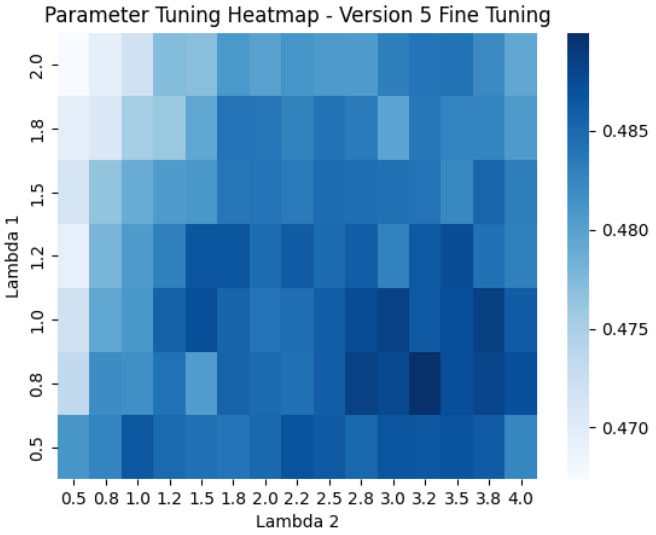

Figure 10. Heatmap of Version 5: Impact of λ_1 and λ_2

Version 5 exhibits a distinct trend where higher $\lambda_2$values combined with lower $\lambda_1$values lead to generally better mAP scores. This could be attributed to the

limited set of source categories available during fine-tuning. Prioritizing pseudolabel learning (high λ_2) helps the model compensate for the reduced source knowledge, maximizing the transfer potential. Minimizing the focus on the seen loss (low λ_1) prevents the model from overfitting to the source categories, potentially enhancing generalizability. Aligning with this finding, we selected the λ_1 as 0.8 and λ_2 as 3.2 in the version 5 model.

5.2.4 Version 6 (Randomly-Split Source Data + Target Data)

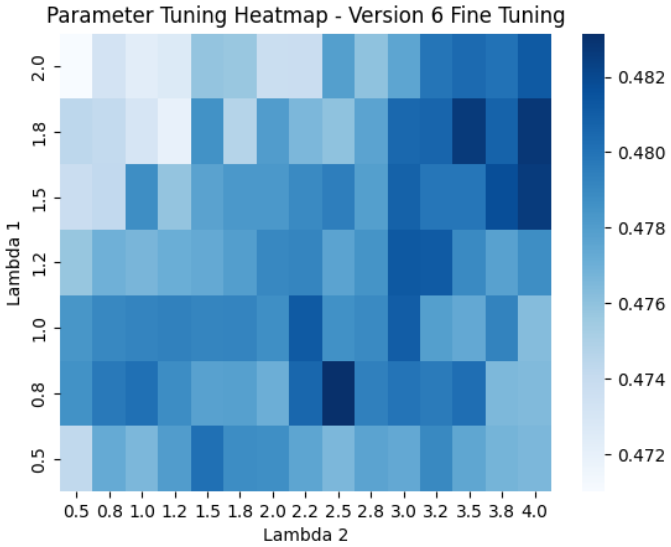

Figure 11. Heatmap of Version 6: Impact of λ_1 and λ_2 on mAP scores

Interestingly, in Version 6, the optimal mAP scores seem to occur when both λ_1 and λ_2 are increased but with few irregularities. The potentially uneven distribution of categories due to random splitting might explain this irregularity. The increased emphasis on both source and target data during training likely forces the model to aggressively learn category representations from the limited source data while simultaneously maximizing knowledge transfer to the target domain. Aligning with this finding, we selected the λ_1 as 0.8 and λ_2 as 2.5 in the version 6 model.

5.2.5 Key Takeaways

- **Prioritizing Target Domain:** Across most versions, increased weighting on λ_2 (unseen loss) is beneficial, underlining the significance of effective pseudolabel learning and knowledge transfer for unlabelled target domains.
- **Contextual Balance:** The optimal balance between λ_1 and λ_2 seems to depend on the quantity and distribution of the source labelled data.
- **Dataset Influence:** The manner in which the source dataset is split has a marked impact on parameter tuning trends.

Chapter 6

Discussion

6.1 Summary of Findings

Our experimental results offer valuable insights into cross-modal retrieval with a focus on knowledge transfer to unlabelled target domains. Key findings include:

1. **Benefits of Joint Learning:** Versions incorporating simultaneous training with both source and target data (as in the original paper and, to some extent, Version 4) demonstrate improved performance. This emphasizes the value of joint learning in guiding knowledge transfer and facilitating adaptation to the target domain, particularly when using pseudolabels.

2. **Importance of Parameter Tuning:** The optimal balance between λ_1 and λ_2 parameters is highly dependent on the specific dataset characteristics. The quantity and distribution of source data categories in both pretraining and fine-tuning stages significantly influence the effectiveness of transfer learning strategies.

3. **Source Data Matters:** Sufficient quantity and diversity within the source labelled data are crucial for both building robust initial representations (pretraining) and enabling effective fine-tuning on the target domain.

6.3 Implications and Future Work

Building upon these findings, several potential strategies can be explored to further optimize adaptation speed and accuracy using joint learning techniques:

1. **Strategic Dataset Splits:** Instead of random source data splits, investigating splits based on category similarity or semantic relatedness could improve knowledge transfer efficiency, potentially influencing optimal parameter choices. This could optimize learning from limited source data.
2. **Improved Pseudolabeling:** Experiment with advanced pseudolabeling techniques to enhance confidence or quality scores associated with pseudolabels. More reliable pseudolabels could further improve knowledge transfer even with limited source data.
3. **Testing on Diverse Datasets:** Evaluating our strategies across a wider range of cross-modal datasets with varying characteristics and diverse categories to assess their broader applicability.

Chapter 7

Conclusion

This thesis investigated the effectiveness of hybrid pre-training and fine-tuning strategies in facilitating knowledge transfer for cross-modal retrieval.

Key findings highlight the advantages of joint learning with source and target data for promoting knowledge transfer. While the two-stage approach of pre-training and fine-tuning demonstrably improved training efficiency in terms of epochs required, the results emphasize the critical role of both source and target domains being present during the same training stage. This simultaneous training allows the model to effectively leverage information from both domains, leading to improved adaptability and performance.

Additionally, the study underscores the importance of dataset-specific parameter tuning. The optimal trade-off between parameters governing the emphasis on source and target data directly depends on the characteristics of the source dataset, including the number and distribution of categories.

Overall, this thesis contributes to the understanding of effective knowledge transfer techniques in cross-modal retrieval. The findings establish a strong foundation for the continued exploration of hybrid training strategies aimed at improving the performance and efficiency of cross-modal retrieval models, while highlighting the importance of considering both source and target data concurrently during training for optimal knowledge transfer.

BIBLIOGRAPHY

K. Wang, Q. Yin, W. Wang, S. Wu, and L. Wang, “A comprehensive survey on cross-modal retrieval,” CoRR, vol. abs/1607.06215, pp. 1–20, 2016. [Online]. Available: http://arxiv.org/abs/1607.06215

L. Zhen, P. Hu, X. Peng, R. S. M. Goh and J. T. Zhou, "Deep Multimodal Transfer Learning for Cross-Modal Retrieval," in IEEE Transactions on Neural Networks and Learning Systems, vol. 33, no. 2, pp. 798-810, Feb. 2022, doi: 10.1109/TNNLS.2020.3029181.

J. C. Pereira et al., “On the role of correlation and abstraction in cross- modal multimedia retrieval,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 36, no. 3, pp. 521–535, Mar. 2014.

N. Rasiwasia et al., “A new approach to cross-modal multimedia retrieval,” in Proc. Int. Conf. Multimedia (MM), 2010, pp. 251–260.

K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” CoRR, vol. abs/1409.1556, pp. 1–14, 2014. [Online]. Available: https://arxiv.org/abs/1409.1556

J. H. Lau and T. Baldwin, “An empirical evaluation of doc2vec with practical insights into document embedding generation,” in Proc. RepL4NLP, 2016, pp. 78–86.

V. Nair and G. E. Hinton, “Rectified linear units improve restricted Boltzmann machines,” in Proc. Int. Conf. Mach. Learn. Madison, WI, USA: Omnipress, 2010, pp. 807–814.

D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” CoRR, vol. abs/1412.6980, pp. 1–15, 2014. [Online]. Available: http://arxiv.org/abs/1412.6980