Meta-Learning For Multi-Modal Cross-Lingual Transfer

Hanxu Hu

Master of Science

Computer Science

School of Informatics

University of Edinburgh

2022

Abstract

Multi-modality, especially vision and language, is an area of great importance in machine learning. Current pre-trained visual linguistic models (PVLMs) have achieved excellent performance in related Vision & Language multi-modal datasets due to the advantage of the pre-training method. Recently, because of the trend of multilingual models, various novel multilingual multi-modal datasets have been built. However, these datasets evaluated current PVLMs and verified that they perform poorly in multi-modal cross-lingual transfer, especially for those low-resource languages. In other words, they show these current PVLMs can't generalize the mapping between texts and images across multiple languages.

To alleviate this problem, we propose a novel meta-learning fine-tuning framework. It makes current PVLMs rapidly adaptive to new languages in multi-modal scenarios. This framework combines one existing cross-lingual meta-learning algorithm with our proposed contrastive meta-learning called Contrastive-MAML. Our proposed framework uses contrastive learning to align the representations that come from different modalities, and uses meta-learning to generalize the alignment to multiple unseen languages. We applied our proposed fine-tuning framework to two different state-of-the-art PVLMs. We verified the effectiveness of our proposed method in two multilingual multi-modal understanding tasks: visual reasoning and visual entailment. The experimental results show that our method can boost the performance of current state-of-the-art PVLMs in both zero-shot and few-shot cross-lingual transfer in these two tasks. We also conduct a series of ablation studies to verify the effect of each component in our method and further verify the rationality of our proposed framework.

Research Ethics Approval

This project was planned in accordance with the Informatics Research Ethics policy. It did not involve any aspects that required approval from the Informatics Research Ethics committee.

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

(*Hanxu Hu*)

Acknowledgements

It has been a challenging and arduous journey, and I would like to thank all those who have helped and supported me. Firstly, I would like to thank my supervisor, Prof. Frank Keller, for his continuous help and guidance and for constructive advice on this project. I would also like to thank my friends, Mr. Tianyi Gao, Mr. Huajian Zhang, and Mr. Xiaoyu Jiang, for their technical and academic help. Finally, I would like to thank my parents for their continuous concern. And I would also like to thank my girlfriend, Miss Qinyi Zhou, who listened to my countless complaints and always comforted me.

Table of Contents

Chapter 1

Introduction

Multi-modal learning is an area of great importance in the field of machine learning. Multi-modal models focus on jointly learning representations from multiple modalities, such as vision and language, so that they can understand the integration information of vision and language. They benefit many downstream tasks, such as Image Caption [42], Natural Language Visual Reasoning [49, 38], Cross-Modal Retrieval [47]. All of these tasks need paired image-text data. Figure 1.1 gives an example of natural language visual reasoning dataset NLVR2 [38]; The input includes two modalities: image and text, and the output is a label which indicates whether the text describes the paired image correctly. Multi-modal learning increases the interaction between different modal information, allowing models to be used in various multimedia applications to enhance the human-computer interaction experience. At the same time, joint learning based on multi-modal data also allows the model to enhance its understanding of a single modality to improve the performance of unimodal tasks, for example, by transferring knowledge of natural language to vision to improve visual performance [35].

Recently, Pre-trained Visual-Linguistic Models (PVLMs) have achieved significant

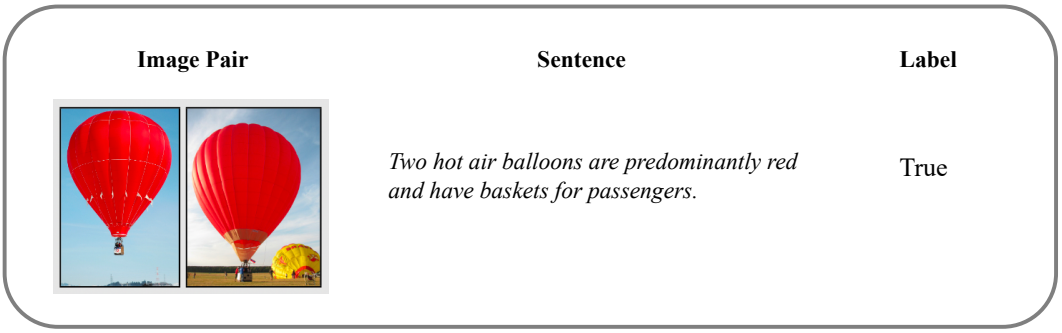


Figure 1.1: Examples in Natural Language Visual Reasoning dataset NLVR2 [38]

success in the multi-modal area. Internet development makes a huge amount of multi-modal data available for pre-training. However, the data which PVLMs learn is mostly in high-resource languages such as English. This gives rise to biases for these models that are extremely dependent on large amounts of training data. Moreover, this bias makes pre-trained multi-modal models perform poorly in the scenario of low-resource languages such as Indonesian and Swahili. As a result, some works about making PVLMs more generalized for multiple languages have been proposed, such as multilingual PVLMs, which are pre-trained by multi-modal data in multiple languages. These models require large amounts of pre-training data in languages other than English, while data in these is hard to collect. Several multilingual multi-modal datasets have been proposed to quantify and evaluate the performance of current multilingual PVLMs in low-resource languages. Studies about these multilingual multi-modal datasets have shown that those models do not perform well in related multilingual downstream tasks, especially in low-resource scenarios such as the setting of zero-shot and few-shot.

Meta-learning can mitigate the issue mentioned above. It is a learning approach that enables machine learning models to adapt quickly to new tasks. It learns the learning algorithm itself, acquiring the ability to learn new tasks quickly. MAML (Model-Agnostic Meta-Learning) [14] is one of the most widely used meta-learning algorithms. It is based on gradient-descent optimization, does not require multiple models or complex settings, and can be used in various models. The detail of MAML is given in Section 2.4. In previous work [41, 14, 31], MAML-based methods have been shown to be useful in low-resource scenarios, such as both few-shot and zero-shot tasks. However, no work has yet attempted to use MAML in multilingual multi-modal tasks to allow models to adapt to new languages quickly.

This thesis is focused on using MAML to address the limitations of previous pre-trained multi-modal models in low resource language multi-modal tasks. We apply a multilingual version of MAML: X-MAML [31] on various PVLMs and validate its effectiveness in zero-shot and few-shot scenarios on multiple multilingual multi-modal datasets. It has been shown that the X-MAML approach can significantly improve the performance of existing PVLMs by fine-tuning them on the auxiliary data in one low-resource language.

We propose a novel unsupervised MAML algorithm using contrastive learning loss as the objective function, called **Contrastive-MAML**. Contrastive learning is an effective and powerful unsupervised learning approach that pairs similar data as positive samples and uncorrelated data pairs as negative samples. It makes the positive samples

closer in the semantic space and the negative samples farther apart, thus allowing for more dispersed and easily discriminable representations of the data. Because multi-modal datasets consist of image-text pairs describing the same or similar objects, labels for downstream tasks are not needed. Based on this, we can take a pair of image and text data from the original dataset as a positive sample and a randomly selected text and image as a negative sample. We use this contrastive learning method as the objective function of MAML, hoping that it can make the model learn how to align and generalize to text and images in new languages.

Finally, we propose a novel meta-learning framework which combines X-MAML and Contrastive-MAML. We follow the X-MAML training approach and use it together with Contrastive-MAML on the training data in one auxiliary language. It can lead to significant improvement for PVLMs in performance in target languages. We also find that using Contrastive-MAML solely in unsupervised scenarios can bring improvements for PVLMs.

We test our proposed framework in **zero-shot** and **few-shot** setups on two multilingual multi-modal datasets: MaRVL [23] and XVNLI [5] for visual reasoning and visual entailment tasks respectively. Zero-shot and few-shot here mean, with no or a few training data in target languages for fine-tuning the model, then directly evaluate the model on test data in target languages. We give a more detailed explanation of zero-shot and few-shot in Section 2.2. Our proposed framework boost the performance of current state-of-the-art pre-trained models in all low-resource languages.

1.1 Contribution

The contributions of this dissertation are summarised below:

- We proposed a general novel meta-learning fine-tuning framework, which could enable models to be quickly adaptive for unseen languages in multi-modal scenarios. Our proposed framework incorporates both supervised and unsupervised learning. It can be used in various multilingual multi-modal tasks, and it can boost the performance of various PVLMs.

- We applied our method to two PVLMs: UC2 [48] and xUNITER [23]. We tested our proposed meta-learning framework in two tasks: multilingual visual reasoning and cross-lingual visual entailment, and verified its effectiveness in both few-shot and zero-shot scenarios.

- We conducted comprehensive ablation studies to verify the effect of each part of our proposed meta-learning framework.

1.2 Outline

Below is the outline describing the remainder of this dissertation:

- **Chapter 2** introduces related works about this thesis, and background knowledge about multi-modal learning, cross-lingual modeling, and multilingual multi-modal representation learning. It also introduces one popular meta-learning method: Model-Agnostic Meta-Learning (MAML).
- **Chapter 3** gives a formal definition of our task and related datasets. It also introduces related evaluation metrics and procedures.
- **Chapter 4** introduces the methodology of this thesis, including the base models we use, the overall framework, and details of each part.
- **Chapter 5** describes the details of our experimental setting and results. Our results show the effect of each part in our proposed framework.
- **Chapter 6** includes our final conclusion, discussion about this conclusion, and possible directions for the future work.

Chapter 2

Background and Related Work

This chapter introduces the necessary background knowledge and recent related works about this thesis. Section 2.1 reviews recent works about multi-modal tasks. Section 2.2 provides recent progress in cross-lingual transfer learning and explains this scenario's zero-shot and few-shot setups. In Section 2.3, we introduce the combination of multi-modal and cross-lingual tasks: multilingual multi-modal representation learning. Then, in section 2.4, we introduce the procedure of the MAML algorithm and its application in the scenario of zero-shot and few-shot settings.

2.1 Multimodal Learning

Multi-modal learning is a scenario of machine learning, where input includes multiple modalities, such as vision and language. The multi-modal models use numerical vectors to represent images and texts, called embeddings. They learn embeddings of each modality and project them into a joint semantic space for representing all modalities together. Then, these models can be used for various multi-modal tasks, such as VQA, cross-modal retrieval, and image description. To formulate this, we consider vision information (image or video) as $\mathbf{v} = \{\mathbf{v}_1, ..., \mathbf{v}_K\}$, and language information (texts) as $\mathbf{w} = \{\mathbf{w}_1, ..., \mathbf{w}_T\}$, where $\mathbf{v}_i$ is the embeddings of i-th region of the image and $\mathbf{w}_j$ is the embeddings of j-th word of the text. The goal of multi-modal models is modelling $P(\mathbf{Y}|\mathbf{v}, \mathbf{w})$, where $\mathbf{Y}$ varies with the specific task. We can decompose this into two part, the first is to learn a joint representation $\mathbf{z}$, which can be formulated as $P(\mathbf{z}|\mathbf{w}, \mathbf{v})$. Then based on the joint representation, we can model the mapping from it to the prediction: $P(\mathbf{Y}|\mathbf{z})$.

There are multiple ways of learning the joint representation. Ngiam et al. [29] firstly

applied a neural network and used autoencoder [17] for learning shared features of multiple modalities in 2011. DeVISE [15] learns visual and textual encoders separately, followed by a linear transformation to the shared space using ranking loss. Wang et al. [43] proposed a two-branch neural network followed by linear and non-linear projections for learning joint embeddings, trained by cross-view constraint and inter-view constraint loss. These methods bridge visual and language information, but all of them are task-specific and cannot be generalized to multiple tasks.

Inspired by BERT [11], which pre-trains transformer-based [40] neural network by the Masking Language Model method [11], some recent works focus on proposing Pre-trained Visual Linguistic Models (PVLMs), and these PVLMs can be used in multiple multi-modal tasks. Pre-training is a training paradigm which trains the model on large-scale data firstly; then, the model can transfer the knowledge they learned into other more specific tasks or smaller datasets. Large-scale pre-training data can make machine learning models gain better initialization parameters than randomly initializing. The architecture of most current pre-trained models is based on transformer [40]. Transformer is an architecture of neural networks which mainly leverages attention mechanism and linear transformation to encode information. Transformer can be stacked into multiple layers and gain a powerful ability to encode sequence or text information. VilBERT [26] and LXMERT [39] use two transformer-based encoders to extract vision and language embeddings independently, then using a third transformer-based encoder to fuse these two modalities. These models can't allow early interaction between modalities, and waste parameters. To tackle this issue, VisualBERT [21], VLBERT [37], and Unicoder-VL [20] use a single transformer-based encoder to extract a joint representation of both image and text, which can enhance interaction between modalities and save space. Based on previous works, UNITER [6] use conditional masking and word region alignment as pre-training tasks to further boost the performance of PVLMs.

Recently, contrastive learning methods have been used in the multi-modal area. Contrastive learning is a set of learning methods. It pairs similar instances as positive examples and dissimilar instances as negative examples, then pushes the distance of negative examples farther and positive examples closer to each other in the representation space of models. Multi-modal area, especially in the situations of vision and language, is very suitable for using contrastive learning because positive image-text pairs examples will be created when constructing datasets. There are already existing works focusing on using contrastive learning for vision and language representation learning. Zhang et al. [46] firstly proposed contrastive learning loss for image-text

pairs in medical images area. CLIP [35] followed the objective function and used it as a pre-training objective function for transferring knowledge from multi-modal information to vision tasks.

Benefiting from large-scale English image-text paired data and powerful pre-training methods, these pre-trained models have the powerful ability to encode image and text information as a joint representation which can be used in multiple downstream multimodal tasks. They achieved great performance in Vision Question Answering [2], Language-Vision Reasoning [38], Image-Text Retrieval, evaluated on related datasets such as NLVR2 [38], RefCOCO [45], and Flickr30K [34]. However, due to the resource limitation, most data for these pre-training tasks only contain English. Limited by their pre-training data, most of their knowledge comes from the Western world and Western culture. This results in their inability to understand diverse cultures and even creates bias.

2.2 Cross-Lingual Modelling

Cross-lingual modeling learns the mapping between languages and projects the embeddings of different languages into the same semantic space. Mikolov et al. [28] used linear transformation to represent the mapping between languages and trained the model to represent embeddings based on a small set of word-level parallel training data. Gouws et al. [16] didn't use word-level paralleled data, but learned multilingual representation by using sentence-level paired bilingual data. These methods are limited by the amount of parallel and bilingual paired data, so unsupervised strategies become essential in cross-lingual scenarios, especially for those low-resource languages which don't have enough amount of parallel training data. To mitigate this problem, Artetxe et al. [3] used self-training method to explore structural similarity of embeddings space, and learn bilingual word embeddings without using bilingua data.

Recently, due to the progress of neural machine translation (NMT) and trasnfer learning, there are some works focusing on transferring knowledge from neural machine translation models to other multilingual tasks. McCann et al. [27] used deep LSTM encoder to contextualize the embeddings of texts, and gain improvements on other common NLP tasks . Eriguchi et al. [13] use the encoder of a multilingual neural machine translation system, followed by a task-specific classifier, which can achieve competitive performance in cross-lingual zero-shot setting (the language of test data is different from training data).

Apart from using the neural machine translation encoder for transfer learning, self-supervised pretraining can also transfer multilingual knowledge from larger datasets to smaller ones about more specific tasks. Devlin et al. [11] released Multilingual BERT (M-BERT), which is pre-trained by monolingual corpora in 104 languages. It achieved excellent performance in cross-lingual zero-shot transfer scenarios. Pires et al. [33] proposed various probing experiments to analyze M-BERT, and verify its ability to capture multilingual representation, especially for typologically similar languages. Conneau et al. [8] proposed two novel pretraining methods for cross-lingual language models (XLMs), which leverage monolingual and parallel data. XLMs used shared vocabulary for tokenization, which improves the alignment between languages in the scenario of shared proper nouns between languages and some special characters. Conneau et al. [7] also proposed XLM-R, which is a multilingual version of RoBERTa [24] (a more robust transformer-based pre-trained model), pre-trained by larger scaled multilingual data. XLM-R has larger parameters than XLM and M-BERT, has a larger vocabulary size, and uses a better sample method across all languages to achieve great performance; in other words, it explores the trade-off of the amount of data across all languages to avoid overfitting to only several languages.

As mentioned above, various pre-training methods make language models effective for multiple downstream tasks. [7, 11] show these pre-trained language models they can also bring improvements even in **zero-shot** and **few-shot** setups. Zero-shot learning is the situation of learning modeling without data in target tasks or classes. And few-shot learning is where there are only a few training data instances in target tasks or classes. In the context of cross-lingual transfer, zero-shot and few-shot settings mean there are no or few training data in target languages. Therefore, in Zero-Shot Cross-Lingual Transfer (ZSCLT) tasks, the convention is to train or fine-tune the models on the training data in source languages (usually English), then evaluate them directly on test data in target languages.

Various downstream tasks and benchmarks can evaluate the performance of cross-lingual transfer methods. Lewis et al. [19] proposed MLQA, a cross-lingual question answering dataset containing 7 languages. Conneau et al. [9] proposed XNLI, which is a cross-lingual version of the visual entailment task. In this task, the system should input two sentences and determine whether they entail, contradicts each other, or neither. It contains 15 languages and even includes several low-resource languages such as Swahili and Urdu. Due to increasing cross-lingual tasks, Hu et al. [18] built XTREME, a benchmark containing multiple cross-lingual datasets for comprehensive

analysis of cross-lingual methods. It integrated several existing datasets, including 40 languages and 9 tasks. It provides an evaluation benchmark of the general performance of multilingual representation learning methods.

2.3 Multilingual Mutimodal Representation Learning

We have introduced multi-modal learning and cross-lingual modeling, while recently, increasing works are trying to investigate the scenario of combining them: multi-modal cross-lingual transfer tasks. Elliott et al. proposed Multi30K [12], an image description dataset which contains multiple languages description. This dataset can be used for both image description and cross-modal retrieval. Pfeiffer et al. [32] built a multilingual version of visual question answering dataset called xGQA. Liu et al. [23] proposed a multilingual version of the grounded visual reasoning dataset called MaRVL, which follow the same setting as NLVR2 [37] (an example of this dataset is in Figure 1.1). Different from previous multilingual multi-modal datasets in which the content across languages is the same, the domain and content of both images and texts of MaRVL are different between languages.

There are several directions to handle these tasks; one is using the pre-training method to build multilingual multi-modal models. Several pre-trained models are recently proposed, focusing on multilingual multi-modal representation learning. Ni et al. [30] proposed M3P, a transformer-based pre-trained model that maps the same concepts in different modalities and languages into a common semantic space. It learns universal representations across modalities and languages. The input includes image embeddings extracted by Faster R-CNN [36], and text embeddings which are the same as BERT [11]. Similar with M3P, Liu et al. [23] extended UNITER [6], proposing mUNITER based on M-BERT [11], and xUNITER based on XLM-R [7]. Chou et al. [48] proposed a data augmented method based on machine translation for cross-lingual cross-modal pre-training, called the pre-trained model UC2. Different from M3P, which uses single multi-layer BERT-like architecture of the neural network, UC2 uses one image encoder to extract information from images based on Faster R-CNN [36] and one cross-lingual language encoder to extract information of language based on XLM-R [7] separately, then uses one cross-lingual cross-modal encoder to merge the feature of two modalities. Although pre-training methods have been proved powerful in multiple tasks, they require extremely large amounts of data for training and have been proven by [23] and [5] to be poor at generalizing for new multilingual multi-modal tasks.

To mitigate above limitations of pre-trained models, adapter method can be a useful way. Adapter method has been widely used for quickly fine-tuning the pre-trained models to adapt to new tasks. The pre-trained model is inserted with extra modular parameters for fine-tuning, while maintaining the pre-trained parameters fixed. Recently, Pfeiffer et al. [32] proposed an adapter method for cross-lingual and cross-modal adapting, and it is effective in multilingual multi-modal tasks. However, this method still introduced additional parameters.

2.4 Model-Agnostic Meta-Learning

Although current multilingual Pre-trained Visual and Language Models can represent embeddings from different languages and modalities in the same semantic space, and can be used in multiple downstream tasks, their generalization ability is not so satisfying. Specifically, there is a significant gap between the performance of these models in English and low-resource languages, especially in zero-shot or few-shot settings. There are some methods try to mitigate this problem, such as adapter method mentioned above, but they still need introducing extra parameters and requiring a lot of training data. To tackle this problem, a better possible solution is Model-Agnostic Meta-Learning (MAML) [14], it is a suitable method for boosting the generalization ability across languages and domains without adding any extra parameters, and only requiring a few data.

MAML is a widely used meta-learning method, which can train a model for quickly adapting to new tasks using a few data.The detail of MAML is mentioned in Algorithm 1. We can denote the distribution of all tasks as $p(\mathcal{T})$, and define a model f_θ with parameters θ. For each task i, we update the original parameters of model θ as $\theta_i^{'}$ by gradient descent method:

$$\theta_i^{'} = \theta - \alpha \nabla_\theta \mathcal{L}_{\mathcal{T}_i}(f_\theta) \tag{2.1}$$

When training, the parameters of model is optimized by all $f_{\theta_i^{'}}$ across tasks $\mathcal{T}_i$ sampled from $p(\mathcal{T})$, which called one meta-optimization. The meta-objective function is :

$$\min_\theta \sum_{\mathcal{T}_i \sim p(\mathcal{T})} \mathcal{L}_{\mathcal{T}_i}(f_{\theta_i^{'}}) = \sum_{\mathcal{T}_i \sim p(\mathcal{T})} \mathcal{L}_{\mathcal{T}_i}(f_{\theta - \alpha \nabla_\theta \mathcal{L}_{\mathcal{T}_i}(f_\theta)}) \tag{2.2}$$

The objective function is the sum of all task losses. Although the loss is calculated by θ_i, it is worth noting that the gradient used for parameter updating in the optimization

Algorithm 1: MAML

input : $p(\mathcal{T})$: distribution over tasks ;
α, β: step size hyperparameters ;
initialized model's parameters θ

while *not done* **do**
  sample batch of tasks $\mathcal{T}_i \sim p(\mathcal{T})$;
  for ***all*** $\mathcal{T}_i$ **do**
    Evaluate $\nabla_\theta \mathcal{L}_{\mathcal{T}_i}(f_\theta)$ with respect to K examples ;
    Compute adapted parameters: $\theta_i^{'} = \theta - \alpha \nabla_\theta \mathcal{L}_{\mathcal{T}_i}(f_\theta)$
  Update $\theta \leftarrow \theta - \beta \nabla_\theta \sum_{\mathcal{T}_i \sim p(\mathcal{T})} \mathcal{L}_{\mathcal{T}_i}(f_{\theta_i'})$

process is calculated based on the derivation of the original θ. Therefore, the optimization process of the MAML algorithm can be regarded as optimizing original parameters by minimizing the losses across several tasks, where the losses are calculated by updated parameters through these tasks, respectively. The purpose of this is to find a better initialization parameter that can quickly adapt to various tasks.

MAML algorithm is widely used in low-resource scenarios such as zero-shot and few-shot tasks. Nooralahzadeh et al. [31] extend MAML for cross-lingual transfer tasks, called X-MAML. They use one low-resource language as an auxiliary language, and the training data contains data both in auxiliary language and English. Each task corresponds to each language. In zero-shot setting, the model is firstly trained by English data, then fine-tuned by X-MAML algorithm on data in auxiliary one language, and finally inference in other low-resource target languages. In few-shot setting, after fine-tuned by X-MAML in auxiliary language, the model will be fine-tuned by a few shots of data in the target language, then tested on the test data in the target language. A more detailed definition and explanation of this algorithm are in Section 4.2. This method has been evaluated in multilingual natural language inference tasks, and achieved significant improvement in zero-shot and few-shot settings.

Chapter 3

Task Definition

In this chapter, we will formally define our tasks, and introduce the detail of datasets we used and their related specific task in Section 3.1. Then we will introduce our evaluation metrics for each task in Section 3.2.

3.1 Datasets and Tasks

Our general task is cross-lingual transfer in vision & language scenarios. We can define the dataset as $\mathcal{X}$. In the dataset, each data point is a pair of visual and language information. The visual information is usually an image or one pair of images, and the language information is a text description corresponding to the visual information. We can represent one data point as $(\mathbf{v}, \mathbf{w})$, where $\mathbf{v}$ is visual information, and $\mathbf{w}$ is language information. The dataset also includes related labels $\mathbf{y}$ for each image-text pair. The label can be varied with the various tasks. For visual grounded reasoning, the label is "True" and "False". For visual entailment task [5, 44], the label is "entailment", "neutral" and " contradiction". For text-image retrieval [22, 34], the label is "relevant" or "irrelevant". So all these three tasks can be regarded as binary or triple classification tasks. For visual question answering, the label is a sequence of text which represent the answer to the question; sometimes, it also can be regarded as a classification task which let the model select the correct answer. Because our task is a cross-lingual transfer task, the language information contains several different languages. For some datasets, one image is paired with several text descriptions in different languages, and texts in low-resource languages are paralleled to texts in English; in this case, the distribution for both vision and language test data in different languages is the same. For others, such as MaRVL [23], different languages have different instances of paired image-text

Figure 3.1: Examples in IGLUE[5] benchmark. The left example comes from MaRVL [23] dataset, and the right example comes from XVNLI dataset proposed in IGLUE.

data, in which case the distribution of both text and images differs between languages.

Recently, Bugliarello et al. [5] built a new benchmark called IGLUE, which integrates four different multi-modal multilingual datasets, including cross-lingual visual entailment (XVNLI), MaRVL [23], xGQA [32], and multilingual cross-modal retrieval (xFlickr & CO). For XVNLI, the image and English text comes from text-only natural language inference task [4], with cross-lingual extension [1] and multi-modal extension [44] . For multilingual cross-modal retrieval, the images comes from the combination of Flickr30K [34] and COCO [22]. Each image is associated with only one English caption, and the captions in other languages are annotated by native speakers.

Because of the limited resource, these datasets only includes data in low-resource languages in validation and test sets. In other words, there is nearly no training data in low-resource languages. So the format of this task is zero-shot or few-shot learning. The multilingual pre-trained visual linguistic models are firstly fine-tuned on data in the source language (usually English), then inference on validation and test set in target languages.

In our project, we use MaRVL [23] and XVNLI [5] datasets to verify the effectiveness of our method. Both of these two tasks can be regarded as a classification problem, and require models have the ability to reason and understand. The examples are shown in Figure 3.1

- **XVNLI** is a cross-lingual visual entailment dataset. The model will get the input of one image and one description, and classify the relationship between the image

and the description from “Entailment”, “Contradiction”, and “Neutral”. The right instance of the Figure 3.1 is an example data point of XVNLI. The given picture describes two people playing American football, but the caption is about basketball, so the output should be “contradiction”. The special feature of this dataset is that it has multilingual text descriptions. It includes four languages other than English: Arabic, French, Spanish, and Russian. Texts from all of these four languages share the same image dataset. The development set and test set are split based on the images to avoid data leaking, and each image corresponds to one text in four languages.

- **MaRVL** is a multilingual visually grounded reasoning dataset. It contains five low-resource languages: Indonesian, Mandarin Chinese, Swahili, Tamil, and Turkish. Each example includes a pair of images, which describes the same concept, and a text description of these two images. The text description includes some reasoning about the related pair of images. It is worth noting that this dataset takes into account the domain shift of data due to different cultural backgrounds. Different from images of English which come from ImageNet [10], the images of low-resource languages come from a different source which collecting by native speakers. Different from previous multilingual datasets, which only consider concepts in the context of the western culture, this dataset also contains some language-specific concepts in data of low-resource languages to mitigate the North American or Western European bias. These language-specific concepts don’t appear in English, but often are found in specific languages. For example, the left instance of the Figure 3.1 shows a data point of MaRVL; both images of this data point describe the same concept “Suona”, one musical instrument in ancient China. It is also worth noting that, in this example, given description includes some numerical and orientation reasoning. What’s more, the same concept in different language backgrounds still might look different. For example, in the second case of Table 5.7, the drum in the left image has a totally different outlook from the drum in the right. The domain shift between languages makes this dataset challenging, requiring the model’s ability to generalization toward languages and domains simultaneously.

3.2 Evaluation Metrics

The evaluation metric depends on the type of specific task. For visual grounded reasoning tasks such as MaRVL [23], there are two metrics, Accuracy and Consistency. Accuracy is the ratio of the number of data that the system predicts correctly to all data. Consistency is the proportion of unique sentences for which the prediction was correct for all corresponding image pairs. For xGQA, the evaluation metric is Accuracy, which is to measure the proportion of questions answered correctly. For image-text retrieval, the evaluation metric is Recall. For XVNLI, the evaluation metric is also Accuracy because it can be regarded as a triple classification problem. Because in this thesis, we only verify our method on XVNLI and MaRVL datasets, the evaluation metric we use is **Accuracy** and **Consistency** for MaRVL, and **Accuracy** for XVNLI. To measure and evaluate the general performance of models, we report the average accuracy or consistency scores across all auxiliary languages to all target languages in both zero-shot and few-shot setups.

Chapter 4

Methodology

This chapter describes the detail of our proposed multi-modal cross-lingual transfer system. We use two multilingual PVLMs: xUNITER [23] and UC2 [48], as our baseline, and introduce them in Section 4.1. In Section 4.2 and 4.3, we describe the detail of two main part of meta-learning framework: X-MAML and Contrastive-MAML separately. Finally, we introduce our overall proposed framework in Section 4.4, which can make current PVLMs more quickly adaptive for multi-modal data in unseen languages.

4.1 Baseline Models

We use multilingual pre-trained visual linguistic models (PVLMs) as our baseline models. As mentioned in Section 2.3, previous Transformer-like pre-trained models have achieved great results in multilingual multi-modal tasks. In our project, we choose xUNITER and UC2 as our baseline models, these two models used different pre-training methods. Then we applied our method on these two models to show that our proposed method is model-agnostic.

4.1.1 xUNITER

xUNITER is a multilingual version of UNITER model [6] proposed by Liu et al. [23]. xUNITER has a similar architecture to UNITER. UNITER's architecture can be present simply as in figure 4.1, where "FC" represents the fully connected layer, and "LN" is the layer normalization. It uses Faster-RCNN [36] as a feature extractor for images. For a single images, 36 objects will be extracted and represented as 36 vectors, we can formulate them as $\mathbf{v} = \{\mathbf{v}_1, ..., \mathbf{v}_M\}$, where $\mathbf{v}$ represent the whole image, and M

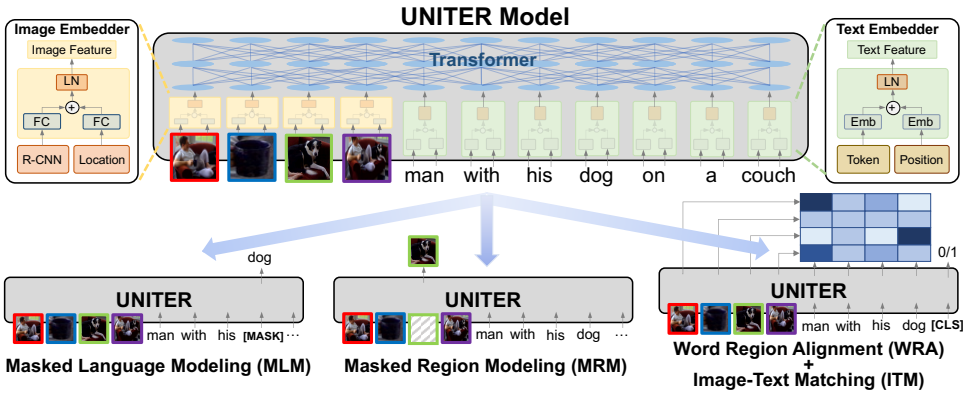


Figure 4.1: Overview of UNITER [6]. (This figure is borrowed from its original paper.)

equals to 36. It is worth noting that the parameters of Faster-RCNN are frozen, which means its parameters will not be updated during training. Then, both features and positional embeddings pass through a fully connected layer, and merge together. The image features are pooled and reshaped as vectors with the same dimension as text embeddings. It is shown that UNITER has four pre-training methods: Masked Language Modelling(MLM), Masked Region Modelling(MRM), Image-Text Matching(ITM), and Word Region Alignment(WRA). For xUNITER, except for the pre-training method mentioned above, it uses Masked Language Modelling for multilingual data, and uses the same text embedder as XLM-R [7] does.

4.1.2 UC2

For UC2, it uses similar model architecture to UNITER, but different pre-training methods. The overview of UC2 is shown in Figure 4.2. The image Encoder is Faster-RCNN [36], and Cross-lingual Language Encoder is similar with the Text Embedder in Figure 4.1. The pre-training method of UC2 augments pre-training English data for constructing multilingual corpus via machine translation, then uses this augmented multilingual data for pre-training. It also proposed Visual Translation Language Modeling(VTLM) pre-training method, which uses the image as a pivot to learn the relationship between parallel texts in two languages and their corresponding images.

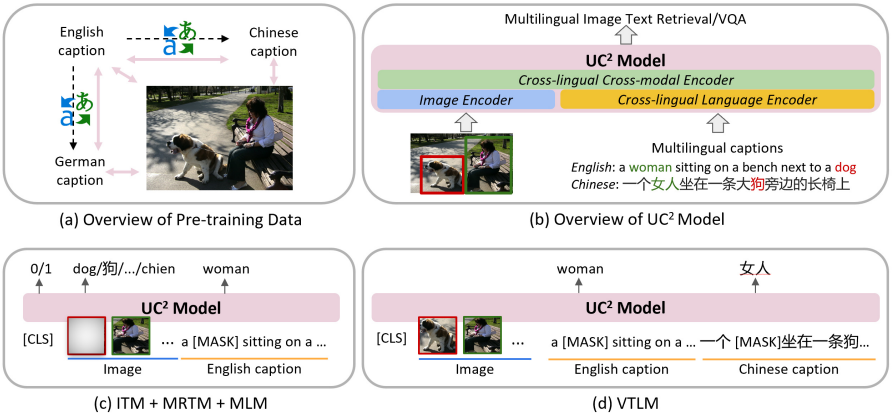


Figure 4.2: Overview of UC2 [48]. (This figure is borrowed from its original paper.)

4.2 X-MAML

We followed the detail of Nooralahzadeh et al. [31] to build the pipeline of X-MAML algorithm, assuming one auxiliary language is annotated and available for training. X-MAML is cross-lingual version of MAML [14], which is a powerful meta-learning method described in Section 2.4. The detailed procedure of X-MAML is shown in Algorithm 2.

Algorithm 2: X-MAML

input : a set of low-resource auxiliary languages (A) ;
pre-trained model's parameters: θ;
α, β step size hyperparameters

while *not done* **do**
  for $l \in A$ **do**
    Sample batch of tasks $\mathcal{T}_i$ using the development set of the language l;
    for *each task* **do**
      Sample K examples for computing $\nabla_\theta \mathcal{L}_{\mathcal{T}_i}(f_\theta)$;
      Compute adapted parameters: $\theta_i' = \theta - \alpha \nabla_\theta \mathcal{L}_{\mathcal{T}_i}(f_\theta)$;
      Sample Q examples for computing $\mathcal{L}_{\mathcal{T}_i}(f_{\theta_i'})$
  Update $\theta \leftarrow \theta - \beta \nabla_\theta \sum_{\mathcal{T}_i \sim p(\mathcal{T})} \mathcal{L}_{\mathcal{T}_i}(f_{\theta_i'})$

We define the set of auxiliary languages as A, and empirical results show that when the size of A is one or two enough for models to get significant improvement which is described in [31]. θ is the parameter of PVLMs, and it is trained by English data

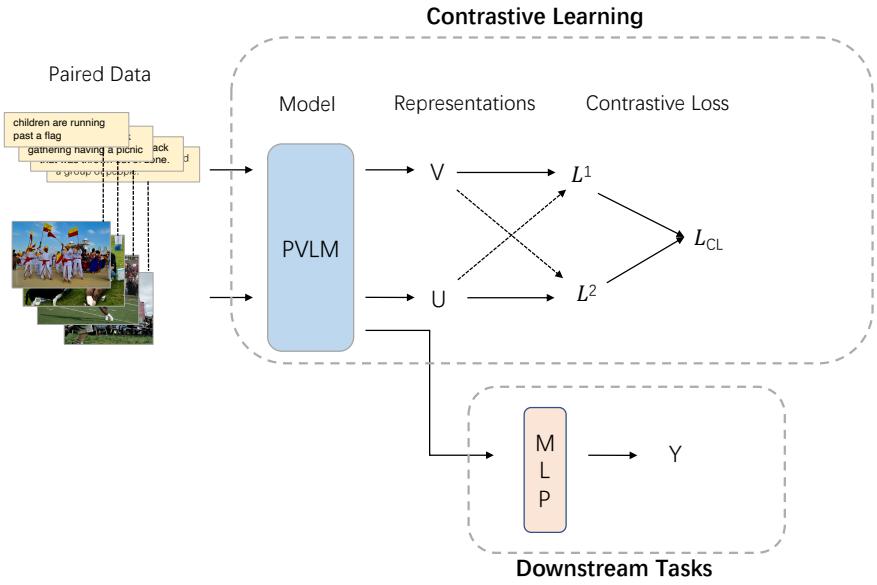


Figure 4.3: The Proposed Architecture of the Model. The architecture consists of a contrastive learning module and downstream tasks module, and both modules share the same parameters of a pre-trained visual language model.

before using X-MAML, just following the procedure in [31]. Then, we sample a batch of tasks $\mathcal{T}_i$ in one auxiliary language. In our specific setting, we define one task as K support examples and Q query examples, and for each language, we only sample one task. The specific optimization procedure is similar to vanilla MAML, which is described in Section 2.4. It is worth noting that each language represents a set of tasks here. Because MAML is for adapting to new tasks, the X-MAML can help models quickly adapt to unseen target languages, which are other than the auxiliary language.

4.3 Contrastive-MAML

Due to the popularity of contrastive learning in multi-modal scenarios, and inspired by X-MAML [31], we try to use contrastive learning loss as the objective function of MAML algorithm. As we described in Section 2.1, contrastive learning is very suitable for the multi-modal area due to positive pairs are created when constructing multi-modal datasets. We followed the multi-modal contrastive learning loss function proposed by Zhang et al. [46], which has been proved effective in medical image scenarios and used as a pre-training objective function by CLIP [35] successfully. We integrated it into our architecture which is shown in Figure 4.3. It can be regarded as an auxiliary task for representation learning, aiming to enable models quickly gain better aligned

multi-modal representation for downstream tasks. We firstly input a batch of image-text pairs, then extract representations of images and texts from the last hidden layer of the Multilingual Pre-trained Visual Linguistic Model as **z**. In the contrastive learning scheme, for texts, we use the representation of the first special token of texts to represent each single text caption. For images, we use the feature of the whole image to represent each image. For a batch of image-text pairs, they can be noted as: $I = \{I_1, ..., I_N\}$, and a batch of texts can be noted as: $T = \{T_1, ..., T_N\}$, where N is the size of batch, and (I_i, T_i) is noted as a image-text pair. In the datasets we used (XVNLI and MARVL), the paired image-text data describe the same or similar concept so we can assume them as positive examples. As Figure 4.3, the representations of images and texts are fed into two different linear transformation layers separately, which are noted as W_1 and W_2.

$$U = I \cdot W_1^\top \tag{4.1}$$

$$V = T \cdot W_2^\top \tag{4.2}$$

Then we can compute the cosine similarity of each pairs as: $\langle U_i, V_j \rangle = \frac{U_i^\top V_j}{\|U_i\| \|V_j\|}$. Our objective is to maximize the similarity of matched image-text pairs, or in other words, let models to predict which pairs of image and text is matched. So the image-text contrastive loss can be formulated as :

$$\mathcal{L}_i^1 = -log \frac{exp(\langle U_i, V_i \rangle)}{\sum_{K=1}^{N} exp(\langle U_i, V_k \rangle)} \tag{4.3}$$

Following the setting in Zhang et al. [46], the contrastive loss should be symmetric for each modality, so we also compute text-image contrastive loss as:

$$\mathcal{L}_i^2 = -log \frac{exp(\langle V_i, U_i \rangle)}{\sum_{K=1}^{N} exp(\langle V_i, U_k \rangle)} \tag{4.4}$$

Finally, we can compute our final contrastive loss of this batch of paired data as:

$$\mathcal{L}_{CL} = \sum_{i=1}^{N} (\mathcal{L}_i^1 + \mathcal{L}_i^2) \tag{4.5}$$

Where $\mathcal{L}_{CL}$ is the overall contrastive loss. When we minimize $\mathcal{L}_{CL}$, we actually maximize the similarity of image-text pairs which are positive examples.

We described the contrastive learning in multi-modal scenarios as above. Then, we will modify this as a meta-learning procedure. Our intuition is that we can use meta-learning such as MAML algorithm, and contrastive loss as its learning objective for fast

adapting multi-modal alignment in new languages. X-MAML [31] algorithm use one auxiliary language as training data and use the loss function of main task as objective function. We use only one auxiliary languages as X-MAML do, but contrastive loss as objective function. Different with X-MAML, our method even don't need annotations of the auxiliary data, but relies only on the paired text-image data itself.

Firstly, we sample a virtual task $\mathcal{T}$, in one auxiliary language which consists of a batch of support data $\mathcal{B}_s$ and a batch of query data $\mathcal{B}_q$. We define the parameters of the model is θ, so we can compute the multi-modal contrastive loss on support data firstly: $L_{CL}(\theta)_{\mathcal{B}_s}$, then the parameters of the model will update by one step of gradient descent:

$$\theta^{'} \leftarrow \theta - \alpha \nabla_{\theta} \mathcal{L}_{CL}(\theta)_{\mathcal{B}_s} \tag{4.6}$$

Following MAML algorithm, our final objective for this task is to minimize $\mathcal{L}_{CL}(\theta^{'})_{\mathcal{B}_q}$ on query data $\mathcal{B}_q$ using gradient descent:

$$\theta \leftarrow \theta - \beta \nabla_{\theta} \mathcal{L}_{CL}(\theta^{'})_{\mathcal{B}_q} = \theta - \beta \nabla_{\theta} \mathcal{L}_{CL}(\theta - \alpha \nabla_{\theta} \mathcal{L}_{CL}(\theta)_{\mathcal{B}_s})_{\mathcal{B}_q} \tag{4.7}$$

Optimized by this method, pre-trained visual-linguistic models can gain the ability of adapting for new tasks in other languages quickly without using any annotations of downstream tasks.

It is worth noting that, just as in Figure 4.3, we regard the Contrastive-MAML algorithm as a procedure of auxiliary task, which will not affect the parameters of Multi-Layer Percetron (MLP) related to downstream tasks. We will describe how this contrastive learning procedure can combine with the training process of downstream tasks in the next section.

4.4 Overall Framework

Both xUNITER and UC2 show competitive performance on multilingual multi-modal datasets when fine-tuned by data in target languages. However, there still is a significant gap between the performance in English and in the low-resource language in **zero-shot** or **few-shot** scenarios, which means if there is no or few data, these pre-trained models will not work well. To mitigate this problem, we proposed a meta-learning fine-tuning framework, combining X-MAML[31] introduced in Section 4.2 and a novel Contrastive-MAML introduced in Section 4.3. We assume data in one low-resource language is available, and use this language as the auxiliary language to help the model learn more generalized representations for cross-lingual multi-modal tasks. This meta-learning

fine-tuning framework is specifically designed for multi-modal cross-language transfer. It can be used in various multi-modal cross-lingual transfer tasks, and it can be applied in various multilingual multi-modal pre-trained models.

Our pipeline of the proposed meta-learning fine-tuning framework can be divided into three parts: (1) fine-tuning the Pre-trained Visual-Linguistic Models (PVLMs) on English counterpart datasets. (2) Using one low-resource language as the auxiliary language and fine-tuning the models on data in this auxiliary language by X-MAML and Contrastive-MAML together. (3) Evaluate the fine-tuned model on the data in target languages. The overall pipeline is shown in Figure 4.4, where θ represents the parameters of PVLM fine-tuned by the English data, which are the initial parameters of the next step: fine-tuning by the combination of Contrastive-MAML and X-MAML on data in one auxiliary language. After this step, the parameters of the final model is θ_{final}, then we will continue zero-shot or few-shot learning on the target languages $t \in \{L \setminus A\}$, where L is the set of all low-resource languages, and A is the auxiliary language. In zero-shot learning, we evaluate the model θ_{final} on the test set in languages t. In few-shot learning, we fine-tune our model θ_{final} by standard supervised learning in a few shots of data in languages t.

Now, we discuss about the detail of combining Contrastive-MAML and X-MAML in our meta-learning framework. The combination will comes true by using the architecture of our model in Figure 4.3, which consists of Contrastive Learning module and Downstream Tasks module. When fine-tuning by Contrastive-MAML and X-MAML together, for each meta-step, we will use a batch of support set $\mathcal{B}_s$ and query set $\mathcal{B}_q$ for both X-MAML and Contrastive-MAML at the same time. For X-MAML, we conduct supervised learning for downstream task, and the procedure is shown in Section 4.2. We denote the loss of downstream task as $\mathcal{L}$, and the temporary parameters optimized for

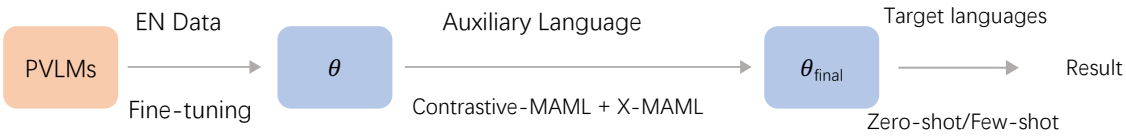


Figure 4.4: The pipeline of our meta-learning framework. The PVLMs are firstly fine-tuned by English Data, which is the same procedure as the baseline. Then using one auxiliary language to conduct our proposed meta-learning fine-tuning framework, which combines the Contrastive-MAML and X-MAML algorithms. Finally, conducting zero-shot or few-shot in target languages to evaluate the performance.

one step by $\mathcal{L}$ on the support set $\mathcal{B}_s$ is $\theta^{''}$. So we can get a gradient $\nabla_\theta \mathcal{L}(\theta^{''})_{\mathcal{B}_q}$. Following the process of Section 4.3, we can also get a gradient $\nabla_\theta \mathcal{L}_{CL}(\theta^{'})_{\mathcal{B}_q})$. Therefore, for combining Contrastive-MAML and X-MAML, we can add these two gradient together when optimizing:

$$\theta \leftarrow \theta - \beta(\nabla_\theta \mathcal{L}(\theta^{''})_{\mathcal{B}_q} + \lambda \nabla_\theta \mathcal{L}_{CL}(\theta^{'})_{\mathcal{B}_q}) \tag{4.8}$$

In the Equation 4.8, β is the meta leanring rate, and λ the scale factor of the Contrastive-MAML. By simply adding the gradients of downstream task and contrastive learning in the procedure of meta-update, we combine X-MAML and Contrastive-MAML in order to gain a better optimization algorithm.

Chapter 5

Experiments and Results

This chapter introduces the experiments setups and related results. For Section 5.1, we describe the details of implementation of our system and settings of experiments. For Section 5.2, we analyze the results we get and the effectiveness of our method in both zero-shot and few-shot scenarios. For Section 5.3, we introduce a series of ablation studies we have conducted and analyze the effect of each parts of our method. For Section 5.4, we conduct case study to show examples of outputs of our system and baseline. For Section 5.5, we discuss the results we gain from our experiments.

5.1 Experiment Setup

We conduct all experiments based on the Visiolinguistic Transformer Architectures framework VOLTA [1]. We implement MAML algorithm based on Higher [2] library. We use AdamW [25] optimizer to fine-tune all models based on PyTorch framework. All experiments are based on PyTorch 1.9.0 and Python 3.7.

5.1.1 Fine-tuning on English Data

Before evaluating models on the data in low-resource languages, we firstly fine-tune the pre-trained models on English datasets, NLVR2 [38] and SNLI-VE [44] respectively for our Visual Reasoning and Visual Entailment tasks following the procedure in [5] and [23]. We follow the setting in IGLUE [5] benchmark: for NLVR2, we set the batchsize as 64, and the size of epochs as 20; for SNLI-VE, we set the batchsize as 128 and the size of epochs as 10. We save parameters of models in each epochs, then picking the

[1] https://github.com/e-bug/volta

[2] https://github.com/facebookresearch/higher

best performance models for each tasks as the initialized parameters θ for following proposed meta-learning fine-tuning stage.

5.1.2 Fine-tuning with Meta-Learning

For X-MAML and Contrastive-MAML algorithms, both the size of support set and query set are 64. We search the learning rates from 5×10^{-5}, 1×10^{-5}, 5×10^{-6} , 1×10^{-6} for both UC2 [48] and xUNIER [23], and find the best learning rate is 5×10^{-6} for both normal fine-tuning stage and meta-update of MAML stage. For the inner learning rate of X-MAML and Contrastive-MAML, we search the learning rate from 5×10^{-6}, 5×10^{-5}, 5×10^{-4} and 5×10^{-3}, and finding that 5×10^{-4} is the best inner learning rate.

For the proposed meta-learning framework, we set the number of iterations as 25, 50, 100, 150, 200, 300, 400, 500, 1000 (for each iterations, we sample a batch of data as support set and a batch as query set) to explore the most suitable number of iterations. We find that models will get overfitting after 300 iterations in most situations, so we set the number of iterations as 400 for all experiments, and evaluate the performance of models for each 25 iterations to guarantee that we can pick the model with best performance of each setting for evaluation.

5.1.3 Zero-shot and Few-shot

We evaluate effectiveness of our proposed method in both zero-shot and few-shot scenarios. For zero-shot, we inference the pre-trained model on test dataset in target languages directly after fine-tuning by proposed meta-learning framework, without seeing any data of taraget languages. For few-shot scenario, after fine-tuning by proposed meta-learning method, we continue fine-tuning the model on several shots of data in target languages for 20 epochs, then save the parameters of the model for each epoch and pick the best one, and evaluate the final model on the test set of target languages. In the final fine-tuning stage in the few-shot setting, we search the learning rate from $\left[1 \times 10^{-5}, 5 \times 10^{-5}, 1 \times 10^{-4}\right]$ just following the setting in IGLUE benchmark [5], and we evaluate our method on 1, 5, 10, 20 shots for XVNLI, and 1, 4, 10 shots for MaRVL respectively.

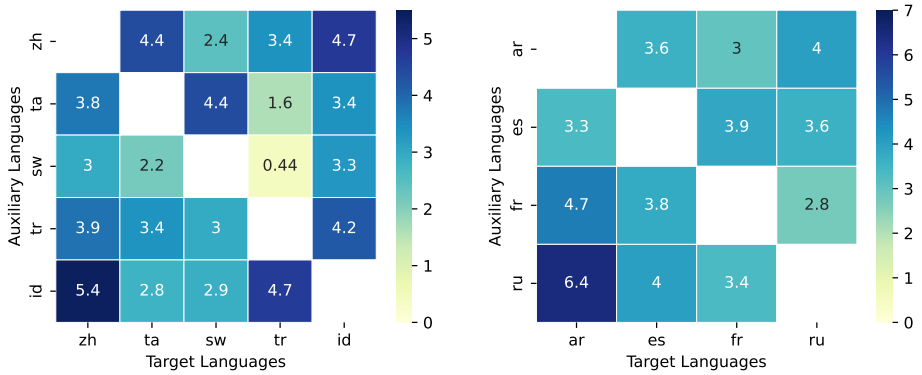


Figure 5.1: Differences in zero-shot performance in terms of accuracy score using xUNITER [23] model fine-tuned by our proposed meta-learning framework. The left heatmap is on MaRVL [23] dataset, and the right is on XVNLI [5] . Rows correspond to auxiliary and columns correspond to target languages.

5.2 Results

In this section, we show the effectiveness of our proposed meta-learning fine-tuning framework by applying it on two state-of-the-art pre-train models UC2 [48] and xUNITER [23] and evaluating on two multilingual multi-modal datasets MaRVL [23] and XVNLI [5]. We assume the labels of one auxiliary language are available, and we evaluate our proposed meta-learning framework for zero-shot and few-shot settings in subsection 5.2.1 and 5.2.2 respectively.

5.2.1 Zero-Shot

We firstly explore the results of the zero-shot setting. We report the difference in the performance of xUNITER fine-tuned with and without using our proposed meta-learning framework in Figure 5.1. It shows the difference in accuracy score for using every single auxiliary language and evaluating data in target languages on both MaRVL and XVNLI datasets. We can observe that there are improvements in performance across all target languages, while the impact across different auxiliary and target languages varies.

We also report the overall results of the baseline models, and they are fine-tuned by our meta-learning framework, which is shown in Table 5.1 and 5.2. It is worth noting that, in our setting, baseline model means PVLMs only fine-tuned on English datasets. For simplification, we only report the average and best performance of using one auxiliary language for each target language. The results for each auxiliary language to

METHOD	ZH	TA	SW	TR	ID	avg
			xUNITER			
Base	54.34/4.74	55.40/6.55	56.41/7.61	57.53/10.99	56.44/7.79	56.02/7.54
Ours ($zh \rightarrow X$)	-	59.82/14.10	58.85/9.78	60.93/13.22	61.17/13.48	-
Ours (AVG)	58.34/9.88	58.49/10.25	59.59/10.33	60.06/12.03	60.35/12.41	59.37/10.98
Ours (MAX)	**59.75/10.28**	**59.82/14.10**	**60.83/10.14**	**62.20/15.25**	**61.17/13.48**	**60.75/12.65**
			UC2			
Base	57.81/12.25	**60.06/11.15**	51.81/1.09	55.76/7.46	56.56/8.51	56.40/8.09
Ours ($zh \rightarrow X$)	-	58.94/12.13	53.61/7.57	55.34/7.99	56.74/8.03	-
Ours (AVG)	58.35/13.44	58.35/12.71	53.99/7.93	56.80/9.61	56.54/9.41	56.81/10.62
Ours (MAX)	**59.59/13.04**	58.94/12.13	**54.60/9.11**	**58.13/13.48**	**56.74/12.60**	**57.60/12.07**

Table 5.1: Zero-shot performance (accuracy/consistency) of two baseline models fine-tuned only by English data (Base) and them finetuned by our proposed meta-learning method (Ours) on the MARVL [23] dataset. Columns indicates low-resource target languages. The avg column indicates the average performance across all low-resource languages in this row. We evaluate pre-trained models on this dataset direclty. We also report results using our proposed method: finetuning on one auxiliary language with X-MAML and Contrastive-MAML. $zh \rightarrow X$ means the auxiliary language is Chinese and the target languages is other low-resource languages X. We also show the average (AVG) and maximum (MAX) performance across all auxiliary languages for each target languages.

each target language are in Table A.1 A.2 A.3. Our proposed meta-learning framework is the combination of Contrastive-MAML and X-MAML. We set the value of λ in the Equation 4.8 as 1×10^{-3} for xUNITER and 0.2 for UC2. While the performance varies across different target languages, our proposed method can boost the performance of UC2 and XUNITER on both MaRVL and XVNLI datasets in general.

Table 5.1 indicates the results on MaRVL dataset. Both pre-trained models UC2 and XUNITER have a poor performance when inference on the low-resource language. The performance is a little different between different target languages. For instance, on the MaRVL dataset, xUNITER baseline model only gets 54.34% accuracy when the target language is Chinese, while getting 57.53% accuracy when the target language is Turkish. In general, both accuracy and consistency improve significantly for both models when using our proposed method. For xUNITER, our method can improve

METHOD	**AR**	**ES**	**FR**	**RU**	**avg**
xUNITER					
Base	53.52	60.05	61.6	61.25	59.10
Ours (ar $\rightarrow$ X)	-	63.66	64.60	65.29	-
Ours (AVG)	58.36	63.86	65.01	64.72	62.99
Ours (MAX)	**59.97**	**64.09**	**64.95**	**65.29**	**63.57**
UC2					
Base	56.70	60.91	68.64	63.91	62.54
Ours (ar $\rightarrow$ X)	-	64.26	68.99	65.72	-
Ours (AVG)	59.94	62.97	69.41	65.18	64.38
Ours (MAX)	**60.65**	**64.26**	**69.73**	**66.07**	**65.18**

Table 5.2: Zero-shot performance (accuracy) of two baseline models only fine-tuned by English data (Base) and them fine-tuned by our proposed meta-learning method (Ours) on XVNLI [5] dataset.

4.73% average accuracy score across all languages when the model used is fine-tuned by the most appropriate auxiliary language (MAX). For UC2, our method also brings improvements from 56.4% to 57.6% for average accuracy performance across all target languages. It is worth noting that our method leads the performance of UC2 on Tamil to decrease. The possible reason is that fine-tuning by auxiliary languages gives the model enhanced generalization performance for all languages, but may decrease the bias for some well-performing languages.

Table 5.2 indicates the results on XVNLI dataset. It gives clear results that our proposed method can perform better than the baseline. In general, these two tables lead to the conclusion that our proposed method can be used in various pre-trained models (at least UC2 and xUNITER), and it is effective in multiple tasks (visual reasoning and visual entailment task). Furthermore, we have observed that, our proposed method can boost the performance of pre-trained models generally, but have less improvement or even worse performance in a few languages. We found that those languages with low improvement or slightly decreasing performance were the languages with higher performance in the baseline model, so we speculate that our method strives to improve the average performance across all languages while improving generalizability, while reducing some of the bias in some languages and reducing the variation in performance

Model	MaRVL		XVNLI
	Accuracy	Consistency	Accuracy
mUNITER [23]	54.01	5.7	53.69
xUNITER [23]	56.02	7.54	59.01
M3P [30]	56.00	-	58.25
UC2 [48]	56.40	8.09	62.54
UC2 (Contrastive-MAML + X-MAML)	56.98	10.62	**64.38**
xUNITER (Contrastive-MAML + X-MAML)	**59.37**	**10.98**	62.99

Table 5.3: The Performances of recent state-of-the-art models in IGLUE [5] benchmark.

between languages.

In summary, we compare the models fine-tuned by our meta-learning framework with the models of previous works about multilingual multi-modal in MaRVL and XVNLI datasets. The result is shown in Table 5.3. The result of mUNITER and M3P comes from IGLUE [5], and the result of xUNITER and UC2 comes from our replication. "UC2(Contrastive-MAML + X-MAML)" and "xUNITER(Contrastive-MAML + X-MAML)" are UC2 and xUNITER fine-tuned by our proposed method, while others are only fine-tuned by English datasets of the same tasks. It indicates the effectiveness of our meta-learning framework, and proves that our method can be generalized to various state-of-the-art pre-trained models and multiple down-stream multilingual multi-modal tasks.

5.2.2 Few-Shot

We also conduct few-shot experiments following the setting in IGLUE [5] for both xUNITER and UC2 on MaRVL and XVNLI. The results are shown in Figure 5.2, where the horizontal axis represents the number of shots, and the vertical axis represents the accuracy score. The leftmost point in the horizontal axis is zero, which represents the performance in zero-shot setups. The red points and lines show the performance of models fine-tuned by our method. The blue lines and points represent the performance of the baseline. It is clear that in all these four figures, our method achieves better performance in all shots. And it is worth noting that although there is a slight increase from the performance of zero-shot to one-shot, our proposed method can still let models

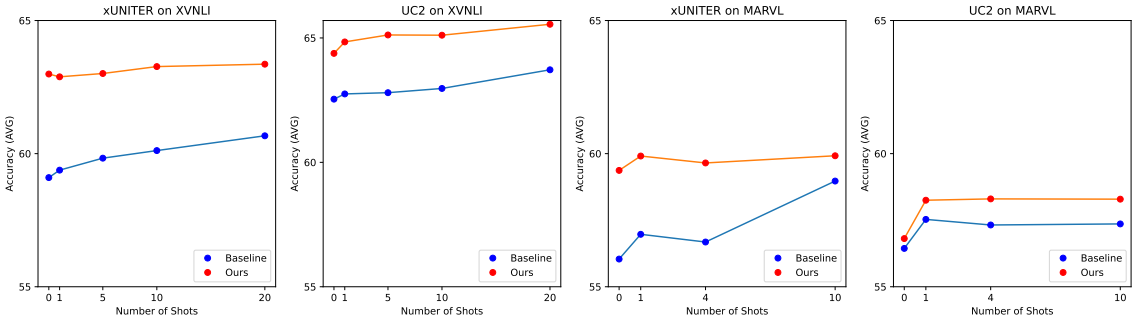


Figure 5.2: Average few-shot performance (accuracy) across all languages of two baseline models on MARVL and UC2 dataset. The horizontal axis represents the number of shots in training data.

gain a better performance without seeing any data in target languages than the baselines seeing a few shots of data in target languages, except for UC2 on MaRVL. In other words, only a few amounts of training data in target languages is not enough to eliminate the difference between baselines and ours caused by the advantage of our method. It further demonstrates that, by using our method, it is only necessary for training data of one auxiliary language, but there is no need for a few training data in each target language.

5.3 Ablation Study

In this section, we conduct a series of ablation studies, which aims to figure out the effect of each part of our proposed meta-learning framework.

- Firstly, we verify the effectiveness of Contrastive-MAML in our meta-learning framework. We compare the performance of using X-MAML solely and combining X-MAML and Contrastive-MAML together, to figure out the effect of our proposed Contrastive-MAML algorithm.

- Secondly, we verify the effectiveness of X-MAML. We further decompose the contribution of X-MAML into introducing the additional training data in one auxiliary language and the MAML algorithm itself. By doing this, we find that, in the X-MAML scheme, only using normal gradient descent fine-tuning in one auxiliary language still can bring significant improvement, but the MAML procedure can further enhance the improvement.

UC2		
Dataset	MaRVL	XVNLI
UC2 (Baseline)	56.40	62.54
UC2 (Contrastive-MAML)	56.82	63.11
UC2 (X-MAML)	56.79	63.81
UC2 (Contrastive-MAML + X-MAML)	**56.98**	**64.38**
xUNITER		
Dataset	MaRVL	XVNLI
xUNITER (Baseline)	56.02	59.11
xUNITER (Contrastive-MAML)	56.84	60.78
xUNITER (X-MAML)	59.23	61.65
xUNITER (Contrastive-MAML + X-MAML)	**59.37**	**62.99**

Table 5.4: Ablation study for verifying the effectiveness of proposed Contrastive-MAML algorithm and X-MAML. This table reports the average accuracy score across all auxiliary languages for all target languages, representing general performance of the model. The full results for different auxiliary languages are shown in Table A.4.

- Finally, we compare our proposed contrastive-MAML algorithm with multi-modal contrastive learning proposed by Zhang et al. [46] which without using MAML, to figure out the rationality of our design for combining MAML with contrastive learning as Contrastive-MAML algorithm.

5.3.1 The Effect of Contrastive-MAML

We verify the effectiveness of the Contrastive-MAML algorithm in our meta-learning framework. Specifically, we compare the performance of the model using both X-MAML and Contrastive-MAML with the model only using X-MAML. The difference between these two models can tell the effect of Contrastive-MAML. The results are shown in the third and fourth rows of Table 5.4. Both UC2 and XUNITER gain a better performance when combining X-MAML and Contrastive-MAML together compared with only using X-MAML. It indicates our Contrastive-MAML algorithm can be used for further boosting the performance of models.

Apart from this, to further explore the effect of the Contrastive-MAML algorithm, we use Contrastive-MAML solely for unsupervised fine-tuning the baseline models. In this setting, we assume there are no labels of downstream tasks available, so we only use unlabelled but paired image-text data in one auxiliary language as training data, and evaluate our fine-tuned model on test data of target languages. The first row reports the average results of the baseline model only fine-tuned by English data, which is the same as the results in Table 5.3, and the second row reports the results of models continually fine-tuned by Contrastive-MAML only using one unlabelled auxiliary data. It demonstrates that Contrastive-MAML can make models achieve better performance in unseen target languages. And Contrastive-MAML has less requirement for auxiliary data than X-MAML (which needs labels of downstream tasks), yet still improves the performance of the model. This unsupervised setting also can be regarded as bringing models the ability to represent data in unseen languages better.

In summary, Table 5.4 illustrate the benefits brought by Contrastive-MAML. It shows that our proposed Contrastive-MAML can be applied solely in an unsupervised learning scenario, and also can be combined with X-MAML in the supervised learning scenario. This further provides evidence that our proposed Contrastive-MAML algorithm can align the representations that come from different modalities and enable models to handle multilingual multi-modal tasks better.

5.3.2 The Effect of X-MAML

In Table 5.4, we can observe the performance of model fine-tuning only by X-MAML from the third row. It brings a significant improvement compared with the performance of baseline. We can also observe that combining the X-MAML and Contrastive-MAML can gain better performance than only using Contrastive-MAML. This further verifies the effect of X-MAML in our meta-learning framework.

As introduced in Section 4.2, the procedure of X-MAML includes using data in one

Model	UC2	xUNITER
Standard Supervised Learning	56.64	59.14
X-MAML	56.79	59.23

Table 5.5: The average accuracy score for standard supervised learning in auxiliary language compared with using X-MAML on MaRVL [23] dataset.

auxiliary language and training by the MAML algorithm. We argue that the contribution of improvement is not entirely due to the algorithm itself because X-MAML also introduces more auxiliary training data. To verify the contribution of training data in auxiliary language and MAML algorithm independently, we fine-tuned UC2 and XUNITER on the MaRVL dataset using the standard gradient descent supervised training method and compared its results with the results of using the X-MAML training method. The results are shown in the first and second rows in Table 5.5. For simplification, all results in this table are the average accuracy scores across all auxiliary languages to target languages. The first row is the average accuracy score of UC2 and xUNITER fine-tuned by standard gradient descent training using data in one auxiliary data. The second row is the average accuracy score of the same two models fine-tuned by the X-MAML algorithm. It indicates that, in the X-MAML training scheme, additional training data brings significant improvement, and the MAML procedure in the X-MAML further enhances the improvement.

5.3.3 The Effect of MAML in Contrastive-MAML

After verifying the effect of Contrastive-MAML, we further compare the performance of our proposed Contrastive-MAML algorithm with the original version of the multi-modal contrastive learning objective proposed by Zhang et al. [46], and try to figure out the effect of MAML algorithm on the contrastive learning. The result is shown in Table 5.6 and Figure 5.3. Although there are one or two languages where Contrastive-MAML does not gain a better result than contrastive learning without using MAML, the result still shows that Contrastive-MAML can achieve better performance in general. This result indicates the rationality of our design which uses the contrastive learning loss as the objective function. We can get the conclusion from our empirical results that regarding contrastive learning as a MAML procedure can lead to the model being more adaptive for unseen languages.

5.4 Case Study

In this section, we visualize some cases of inputs and related predictions of baseline models and the models fine-tuned by our method. In Table 5.7, we use xUNITER to predict the Chinese part of the MaRVL dataset. We have selected three examples where baseline predicted wrongly but our method predicted correctly, and two examples where

Model	xUNITER	UC2
Baseline	56.02	56.4
Contrastive Learning [46]	56.66	56.44
Contrastive-MAML	**56.84**	**56.82**

Table 5.6: Comparison between Contrastive-MAML and multi-modal contrastive learning proposed by Zhang et al. [46] on MaRVL [23] dataset.

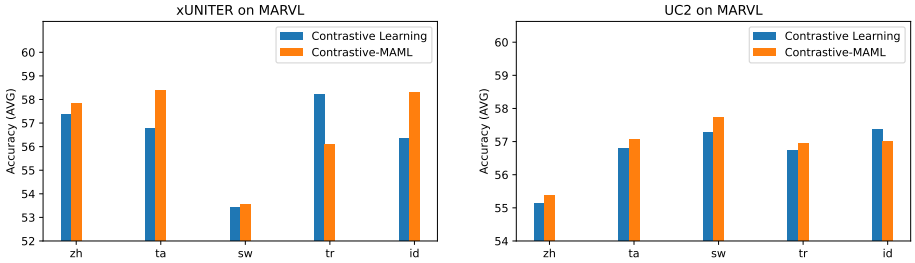


Figure 5.3: The impact of Contrastive-MAML and Multi-modal Contrastive Learning [46] when different languages as auxiliary languages.

both our method and baseline method predicted correctly. We can see that in the first three examples, the label was True but the baseline predicted False. We find that the same concepts have different visual features in the left and the right image for each example, which makes it more difficult for models to identify. For instance, in the first example, the dining rooms in the left and right images look different, and in the right images, the dining room apparently has a more Chinese style. In the second example, the drums in the left and right images have entirely different shapes and colors. In the last two examples, however, the concepts described in the text do not have diverse or obscure visual features when they appear in the images. On the contrary, both panda and rose in these images are easily recognizable. Therefore, based on these cases, we can speculate the meta-learning framework makes the model more adaptive for diverse information, and have better generalization capabilities of diverse mapping between texts and images.

5.5 Discussion

In this thesis, we propose a meta-learning framework to improve the model's generalization ability across languages. Experimental results show that both X-MAML and

Contrastive-MAML can bring improvements in this multi-modal cross-lingual transfer situation. When exploring the effects of each part in our proposed meta-learning framework, we also observe that introducing auxiliary language as extra training data is very helpful. Although there are gains in performance for all auxiliary languages, which has shown in Figure 5.1, the gains are different across languages. This indicates the performance is dependent on the choice of auxiliary languages. One possible hypothesis of the various gains in performance across different auxiliary languages is the typological correlations between auxiliary and target languages, which has been analyzed in Nooralahzadeh et al. [31]. But different from [31], which didn't gain improvement when using Swahili as an auxiliary language in cross-lingual transfer scenarios, we gained significant improvement when using Swahili in our tasks. One possible reason is that, except for natural language, we also have image information in our tasks, which can also provide extra knowledge to help model learning. In general, the choice of different auxiliary languages and the impact of image data on the target language in multi-modal scenarios is still worth further investigating.

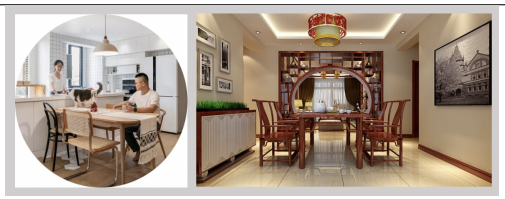
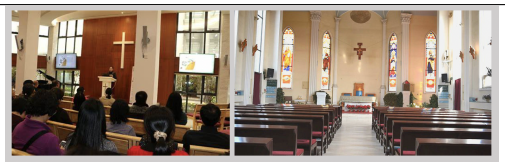
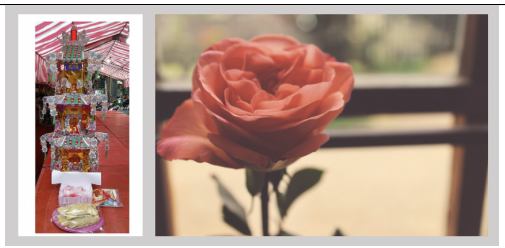
Images	Label	Pred(Base)	Pred(Ours)
 左圖中飯廳恰好有兩個人，右圖的飯廳則沒有人。 (The dining room on the left is occupied by exactly two people, while the dining room on the right is empty.)	True	False	True
 左圖可以看到有八面或以上的鼓，右圖中的鼓面則是金色或咖啡色。 (On the left, you can see a drum with eight sides or more, and the drum on the right is gold or brown.)	True	False	False
 兩張圖都是教堂內部的照片。 (Both pictures are photos of the interior of the church.)	True	False	True
 其中一張圖中有明顯且完整的玫瑰花，另一張圖則沒有。 (There are obvious and complete roses in one picture, but not in the other picture.)	True	True	True
 兩張圖中都恰好只有一隻熊貓，而且都沒有在移動。 (There is exactly one panda in both pictures, and neither of them is moving.)	False	False	False

Table 5.7: Case Study. The red text represents a correct prediction (same as the label). The green text represents a wrong prediction (different from the label). " Pred(Base) " represents the predictions of the baseline of xUNITER. "Pred(Ours)" represents the predictions of xUNITER using our method.

Chapter 6

Conclusion and Future Work

6.1 Conclusion

In this thesis, we focus on mitigating the problem of poor performance of current pre-trained visual linguistic models (PVLMs) in multi-modal cross-lingual transfer scenarios. Most of the pre-training data of PVLMs comes from existing English datasets, or translated datasets, which make PVLMs biased toward the English Language, even concepts in English or Western culture. Based on this, we use meta-learning to make pre-trained models quickly adaptive for new languages in multi-modal scenarios, which no one has explored before. Specifically, we design a meta-learning framework for handling this problem. The reason is that the nature of meta-learning is letting models quickly adapt to new tasks, where we can regard new languages as new tasks in this scenario.

Our meta-learning framework consists of two components, X-MAML and our proposed Contrastive-MAML. X-MAML is a cross-lingual version of MAML algorithm, which need data in one auxiliary language for fine-tuning the model. We implement and verify the effectiveness of X-MAML algorithm in the multi-modal cross-lingual transfer scenario. Contrastive-MAML is a novel training algorithm that combines MAML and contrastive learning. We design the algorithm like this because of its nature for multi-modal and cross-lingual scenarios: the MAML algorithm can make the model adaptive for unseen languages, and contrastive learning can align representations from different modalities. Finally, we combine X-MAML and Contrastive-MAML by adding their gradients during the meta-update step in the MAML procedure.

Experimental results demonstrate that our proposed meta-learning framework can significantly improve the performance of models in multi-modal cross-lingual transfer

both in zero-shot and few-shot setups. We applied our method to two state-of-the-art PVLMs, UC2 and xUNITER, and verified the effectiveness on MaRVL and XVNLI datasets. We also conducted a series of ablation studies to explore the effect of each part of our proposed meta-learning framework. The results of the ablation study show that both X-MAML and Contrastive-MAML bring a positive impact. Furthermore, our proposed Contrastive-MAML can be used solely in the unsupervised scenario without needing any labeled data, and can significantly boost the performance of PVLMs. We also compare the performance of using Contrastive-MAML with contrastive learning without MAML, and it demonstrates the superiority and rationality of our proposed Contrastive-MAML. Our experimental results also suggest that auxiliary language is important in this framework. We discuss possible patterns and reasons for differences in performance when using different auxiliary languages.

6.2 Future Work

Although having proposed a novel method and conducted a series of analyses, we still have limitations that should be solved in the future. Firstly, we only conducted experiments on two multilingual multi-modal datasets, MaRVL and XVNLI, and both are about natural language understanding. We need to apply our method on more various tasks, such as cross-modal retrieval and visual question answering, to further verify our proposed method. Especially for cross-modal retrieval, because our proposed Contrastive-MAML makes instances in similar semantics but different modalities closer, which is similar to the nature of retrieval tasks: finding related information.

Secondly, our proposed Contrastive-MAML can be further optimized. Our proposed Contrastive-MAML can only be seen as an alignment between different modalities at an instance level. But we can further explore more fine-grained alignment, such as at the object level. For instance, we can align the words from texts with objects from images, rather than the whole sentence with the whole image.

Thirdly, we have discussed possible reasons for the difference in the performance of using different auxiliary languages, but we still don't have a quantified evaluation and exploration of this. It is worth figuring out the exact correlation and pattern between auxiliary and target languages. It is also worth using two or more languages as auxiliary languages to conduct experiments.

Bibliography

[1] Željko Agić and Natalie Schluter. Baselines and test data for cross-lingual inference. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 2018.

[2] 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, 2015.

[3] Mikel Artetxe, Gorka Labaka, and Eneko Agirre. Learning bilingual word embeddings with (almost) no bilingual data. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 451–462, 2017.

[4] Samuel Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, 2015.

[5] Emanuele Bugliarello, Fangyu Liu, Jonas Pfeiffer, Siva Reddy, Desmond Elliott, Edoardo Maria Ponti, and Ivan Vulić. Iglue: A benchmark for transfer learning across modalities, tasks, and languages. *arXiv preprint arXiv:2201.11732*, 2022.

[6] 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, 2020.

[7] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Édouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale.

In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, 2020.

[8] Alexis Conneau and Guillaume Lample. Cross-lingual language model pretraining. *Advances in neural information processing systems*, 32, 2019.

[9] Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. Xnli: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, 2018.

[10] 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, 2009.

[11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, 2019.

[12] Desmond Elliott, Stella Frank, Khalil Sima'an, and Lucia Specia. Multi30k: Multilingual english-german image descriptions. *arXiv preprint arXiv:1605.00459*, 2016.

[13] Akiko Eriguchi, Melvin Johnson, Orhan Firat, Hideto Kazawa, and Wolfgang Macherey. Zero-shot cross-lingual classification using multilingual neural machine translation. *arXiv preprint arXiv:1809.04686*, 2018.

[14] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*, pages 1126–1135. PMLR, 2017.

[15] 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.

[16] Stephan Gouws, Yoshua Bengio, and Greg Corrado. Bilbowa: Fast bilingual distributed representations without word alignments. In *International Conference on Machine Learning*, pages 748–756. PMLR, 2015.

[17] Keith J Holyoak. Parallel distributed processing: explorations in the microstructure of cognition. *Science*, 236:992–997, 1987.

[18] Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *International Conference on Machine Learning*, pages 4411–4421. PMLR, 2020.

[19] Patrick Lewis, Barlas Oguz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. Mlqa: Evaluating cross-lingual extractive question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7315–7330, 2020.

[20] 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, 2020.

[21] 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.

[22] 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 *European conference on computer vision*, pages 740–755. Springer, 2014.

[23] Fangyu Liu, Emanuele Bugliarello, Edoardo Maria Ponti, Siva Reddy, Nigel Collier, and Desmond Elliott. Visually grounded reasoning across languages and cultures. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10467–10485, 2021.

[24] 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.

[25] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2018.

[26] 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.

[27] Bryan McCann, James Bradbury, Caiming Xiong, and Richard Socher. Learned in translation: Contextualized word vectors. *Advances in neural information processing systems*, 30, 2017.

[28] Tomas Mikolov, Quoc V Le, and Ilya Sutskever. Exploiting similarities among languages for machine translation. *arXiv preprint arXiv:1309.4168*, 2013.

[29] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y Ng. Multimodal deep learning. 2011.

[30] Minheng Ni, Haoyang Huang, Lin Su, Edward Cui, Taroon Bharti, Lijuan Wang, Dongdong Zhang, and Nan Duan. M3p: Learning universal representations via multitask multilingual multimodal pre-training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3977–3986, 2021.

[31] Farhad Nooralahzadeh, Giannis Bekoulis, Johannes Bjerva, and Isabelle Augenstein. Zero-shot cross-lingual transfer with meta learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4547–4562, 2020.

[32] Jonas Pfeiffer, Gregor Geigle, Aishwarya Kamath, Jan-Martin Steitz, Stefan Roth, Ivan Vulić, and Iryna Gurevych. xgqa: Cross-lingual visual question answering. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2497–2511, 2022.

[33] Telmo Pires, Eva Schlinger, and Dan Garrette. How multilingual is multilingual bert? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, 2019.

[34] 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, 2015.

[35] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark,

et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021.

[36] 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.

[37] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. Vl-bert: Pre-training of generic visual-linguistic representations. In *International Conference on Learning Representations*, 2019.

[38] Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. A corpus for reasoning about natural language grounded in photographs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6418–6428, 2019.

[39] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*, 2019.

[40] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[41] Vinay Kumar Verma, Dhanajit Brahma, and Piyush Rai. Meta-learning for generalized zero-shot learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 6062–6069, 2020.

[42] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164, 2015.

[43] Liwei Wang, Yin Li, and Svetlana Lazebnik. Learning deep structure-preserving image-text embeddings. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5005–5013, 2016.

[44] Ning Xie, Farley Lai, Derek Doran, and Asim Kadav. Visual entailment: A novel task for fine-grained image understanding. *arXiv preprint arXiv:1901.06706*, 2019.

[45] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *European Conference on Computer Vision*, pages 69–85. Springer, 2016.

[46] Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D Manning, and Curtis P Langlotz. Contrastive learning of medical visual representations from paired images and text. *arXiv preprint arXiv:2010.00747*, 2020.

[47] Liangli Zhen, Peng Hu, Xu Wang, and Dezhong Peng. Deep supervised cross-modal retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10394–10403, 2019.

[48] Mingyang Zhou, Luowei Zhou, Shuohang Wang, Yu Cheng, Linjie Li, Zhou Yu, and Jingjing Liu. Uc2: Universal cross-lingual cross-modal vision-and-language pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4155–4165, 2021.

[49] Stephanie Zhou, Alane Suhr, and Yoav Artzi. Visual reasoning with natural language. *arXiv preprint arXiv:1710.00453*, 2017.

Appendix A

Full results

AUXILIARY/TARGET	AR	ES	FR	RU
xUNITER on XVNLI				
AR	-	63.66	64.61	65.29
ES	56.87	-	65.46	64.86
FR	58.25	63.83	-	64
RU	59.97	64.09	64.95	-
UC2 on XVNLI				
AR	-	64.26	68.99	65.72
ES	60.481	-	69.729	66.06
FR	58.68	61.512	-	63.75
RU	60.65	63.14	69.51	-

Table A.1: Full **accuracy** scores of our method on XVNLI [5] dataset when each languages as auxiliary language and target language. Rows corresponds to auxiliary languages, and columns corresponds to target languages

AUXILIARY/TARGET	ZH	TA	SW	TR	ID
xUNITER on MaRVL					
ZH	-	59.823	58.845	60.932	61.17
TA	58.103	-	60.83	59.153	59.84
SW	57.312	57.568	-	57.966	59.752
TR	58.202	58.776	59.386	-	60.638
ID	59.746	58.202	59.296	62.203	-
UC2 on MaRVL					
ZH	-	58.937	53.61	55.339	56.738
TA	57.016	-	53.52	57.034	55.94
SW	59.19	58.776	-	58.136	56.738
TR	59.585	58.535	54.242	-	56.738
ID	57.213	57.085	54.603	56.695	-

Table A.2: Full **accuracy** scores of our method on MaRVL [23] dataset when each language as auxiliary language and target language. Rows corresponds to auxiliary languages, and columns corresponds to target languages

AUXILIARY/TARGET	ZH	TA	SW	TR	ID
xUNITER on MaRVL					
ZH	-	14.098	9.783	13.22	13.475
TA	9.091	-	10.140	10.169	13.121
SW	9.091	7.541	-	9.492	10.284
TR	11.067	10.164	11.232	-	12.765
ID	10.277	9.180	10.145	15.254	-
UC2 on MaRVL					
ZH	-	12.131	7.573	7.991	8.031
TA	10.672	-	6.049	7.039	7.505
SW	16.206	13.978	-	13.482	12.599
TR	13.043	12.366	8.993	-	9.498
ID	13.834	12.366	9.113	9.920	-

Table A.3: Full **consistency** scores of our method on MaRVL [23] dataset when each languages as auxiliary language and target language. Rows corresponds to auxiliary languages, and columns corresponds to target languages

METHOD/TARGET LANG	ZH	TA	SW	TR	ID
UC2					
Baseline	57.81/12.25	60.06/11.15	51.81/1.09	55.76/7.46	56.56/8.51
Contrastive Learning	57.39/10.87	56.78/10.58	53.45/8.29	58.24/9.82	56.34/9.41
Contrastive-MAML	57.83/11.56	58.39/10.03	53.57/7.78	56.12/8.82	58.29/9.75
X-MAML	58.52/12.55	58.37/11.87	52.73/7.69	56.48/8.50	56.71/8.99
Contrastive-MAML+X-MAML	58.25/13.44	58.33/12.71	53.99/7.93	56.80/9.61	57.54/9.41
xUNITER					
Baseline	54.34/4.74	55.40/6.55	56.41/7.61	57.53/10.99	56.44/7.79
Contrastive Learning	55.14/5.23	56.8/7.11	57.29/8.12	56.74/7.73	57.36/8.37
Contrastive-MAML	55.36/5.53	57.06/7.23	57.74/7.65	56.95/7.51	57.00/8.18
X-MAML	58.5/10.66	58.81/10.55	59.74/11.33	59.47/11.27	59.64/12.23
Contrastive-MAML+X-MAML	58.34/9.88	58.74/10.25	59.59/10.33	60.06/12.03	60.35/12.41

Table A.4: Average accuracy scores across all auxiliary languages of ablation studies on MaRVL [23] dataset for each target language.