Meta-learning For Vision-and-language Cross-lingual Transfer

Hanxu Hu
University of Edinburgh
School of Informatics
hanxu.hu@ed.ac.uk

Frank Keller
University of Edinburgh
School of Informatics
keller@inf.ed.ac.uk

Abstract

Current pre-trained vision-language models (PVLMs) achieve excellent performance on a range of multi-modal datasets. Recent work aims at building multilingual versions of such models, and a range of multilingual multi-modal datasets have been introduced for this purpose. However, current PVLMs typically perform poorly on such datasets when used for zero-shot or few-shot cross-lingual transfer, especially for low-resource languages. To alleviate this problem, we propose a novel meta-learning fine-tuning framework. Our framework makes it possible to rapidly adapt PVLMs to new languages by using Model-agnostic Meta-learning (MAML) in a novel cross-lingual multi-modal manner. Experiments show that this new method boosts the performance of current PVLMs in both zero-shot and few-shot settings on four different vision-language tasks across 14 languages.

1 Introduction

Multi-modal models focus on jointly learning representations from multiple modalities, such as vision and language. Many tasks require the integration of information of vision and language, including image captioning (Vinyals et al., 2015), natural language visual reasoning (Zhou et al., 2017; Suhr et al., 2019), and cross-modal retrieval (Zhen et al., 2019). Multi-modal learning captures the interaction between different modalities, allowing the resulting representations to be used in multimedia applications that enhance human-computer interaction.

Recently, pre-trained vision-language models (PVLMs; Chen et al. 2020; Lu et al. 2019; Tan and Bansal 2019) have achieved significant advances in multi-modal tasks. However, the data which PVLMs learn from is mostly for high-resource languages such as English. The resulting models rely on large amounts of training data for good performance, and often the models acquire biases

Figure 1: Examples in IGLUE (Bugliarello et al., 2022) benchmark. The left example comes from MaRVL (Liu et al., 2021) dataset, and the right example comes from XVNLI dataset proposed in IGLUE.

that mean they perform poorly in low-resource languages such as Indonesian or Swahili. To address this, several multilingual PVLMs have been proposed (Zhou et al., 2021; Ni et al., 2021). A number of studies have used multilingual multi-modal datasets (Bugliarello et al., 2022; Liu et al., 2021) and Figure 1 shows two examples from such datasets. The authors of these datasets used them to evaluate current famous PVLMs and demonstrated they do not perform well in low-resource cross-lingual transfer settings.

In this paper, we conjecture that meta-learning can mitigate this issue. This is a learning approach that enables machine learning models to adapt quickly to new tasks by learning the learning algorithm itself. Model-agnostic Meta-learning (MAML; Finn et al. 2017) is one of the most widely used meta-learning frameworks. It is based on gradient-descent optimization, does not require multiple models or complex settings, and can be used for a range of models. In previous work (Verma et al., 2020; Finn et al., 2017; Nooralahzadeh et al., 2020), MAML-based methods have been shown to be useful in low-resource and cross-lingual transfer scenarios, including both few-shot and zero-shot cross-lingual tasks. However, prior work has only attempted to use MAML for cross-lingual transfer in **text-only tasks**

(Nooralahzadeh et al., 2020).

Inspired by previous works about using MAML for natural language tasks, this paper proposes XVL-MAML, a novel variant of MAML that addresses the limitations of previous PVLMs in **vision-language tasks** for low-resource cross-lingual transfer. Our framework combines a traditional supervised loss for learning down-stream tasks with a contrastive loss to encourage the alignment between modalities, resulting in a cross-lingual, multi-modal MAML optimization procedure.

The intuition underlying our method is that a contrastive loss can align representations of different modalities, and MAML allows the model to generalize quickly to unseen tasks (languages, in our case). We show that XVL-MAML can lead to significant improvements in PVLM performance for cross-lingual transfer. We also find that using contrastive learning in a MAML framework on its own can bring improvements in PVLM performance in unsupervised settings.

In sum, our contributions are as follows: (1) We propose a novel framework called XVL-MAML which is the first meta-learning method specialized for vision-language cross-lingual transfer, and doesn't require the translation or pre-training data. (2) We show that using only contrastive learning in the MAML framework in an unsupervised setting can also be useful. (3) We demonstrate that our proposed framework can boost the performance of current PVLMs across 14 languages and four tasks in both **zero-shot learning** and **few-shot learning**. (4) We conduct an ablation study to verify the effect of contrastive learning in both supervised and unsupervised settings and present an analysis across languages and tasks.

2 Related Work

2.1 Multilingual Vision-and-Language Methods and Tasks

Recent work has investigated vision-and-language cross-lingual transfer tasks. Elliott et al. (2016) proposed Multi30K, an image description dataset which contains descriptions in multiple languages. Previous methods (Gella et al., 2017; Rotman et al., 2018) propose ways of bridging languages through images, but they mainly focus on image-text retrieval and only consider high-resource languages such as English and German. Pfeiffer et al. (2022) built a multilingual visual question answer-

ing dataset xGQA. Liu et al. (2021) proposed a multilingual version of the grounded visual reasoning dataset MaRVL, which follow the same setting as the natural language visual reasoning dataset NLVR2 (Su et al., 2019), but considers both cross-lingual transfer and domain shift between languages.

Several pre-trained models are recently proposed for vision-and-language cross-lingual transfer. Ni et al. (2021) proposed M3P, a transformer-based pre-trained model that maps the same concepts in different modalities and languages into a common semantic space. Similar to M3P, Liu et al. (2021) extended UNITER (Chen et al., 2020), proposing mUNITER based on M-BERT (Devlin et al., 2019), and xUNITER based on XLM-R (Conneau et al., 2020). Zhou et al. (2021) proposed UC2, a model using a data augmentation method based on machine translation for cross-lingual cross-modal pre-training. Although pre-training methods have proven powerful across multiple tasks, they require large amounts of training data and show a clear performance gap between English and low-resource languages on the IGLUE benchmark (Bugliarello et al., 2022).

Recently, some adapter-based efficient tuning methods (Pfeiffer et al., 2022; Wang et al., 2023) and translation augmented methods (Qiu et al., 2022) were proposed for multilingual multimodal tasks. But they still require a large amount of data or machine translated data for training. Our method, in contrast, only requires a small amount of auxiliary data.

2.2 Meta-Learning

Meta-learning has been increasingly popular in machine learning. Whereas conventional machine learning methods learn by data points, meta-learning learns by tasks. Previous meta-learning work (Vinyals et al., 2016; Finn et al., 2017) focused on adapting to new tasks quickly. But meta-learning can be applied to other scenarios as well, including semi-supervised learning (Ren et al., 2018), multi-task learning (Yu et al., 2020), and domain generalization (Li et al., 2018).

Prior work has also explored the effectiveness of meta-learning in NLP: Wang et al. (2021) applied meta-learning in semantic parsing for domain generalization based on MAML (Finn et al., 2017; Li et al., 2018). Obamuyide and Vlachos (2019) leveraged meta-learning under limited su-

pervision in a relation classification task. Recently, there have been some applications using MAML in cross-lingual transfer: Gu et al. (2018) and Nooralahzadeh et al. (2020) regard languages as tasks in their meta-learning framework. In contrast to these existing approaches, which explore text-only scenarios, we are the first to utilize meta-learning for cross-lingual transfer in multi-modal tasks.

3 Meta-learning for Vision-and-language Cross-lingual Transfer

We first formally define the problem of vision-and-Language cross-lingual transfer in the context of zero-shot and few-shot scenarios in Section 3.1. Then, we introduce our overall fine-tuning framework in Section 3.2. And we introduce the contrastive learning used for vision-and-language tasks in Section 3.3. Finally, we introduce our XVL-MAML algorithm in Section 3.4.

3.1 Problem Definition

Following the multilingual vision-language IGLUE benchmark (Bugliarello et al., 2022), we formulate the problem of cross-lingual transfer learning in vision-and-language scenarios. For understanding tasks, the input is a pair of an image V and text U, and the output Y is the result inferred by the multi-modal model. We can thus formulate this problem as computing $P_\theta(Y|V, U)$, where θ are the parameters of the PVLMs. During training, the image-text pairs come from datasets D_s in a set of source languages, and our aim is to perform well on datasets D_t for the same task in the target languages. For the zero-shot setup, the pre-trained model fine-tuned on D_s is directly used in inference on D_t for unseen target languages. For the few-shot setup, after training on D_s, the model is continually fine-tuned on several shots of the training set of D_t and then evaluated on the development set of D_t.

3.2 Overall Fine-tuning Framework For Cross-lingual Transfer

The pipeline of our proposed meta-learning fine-tuning framework can be divided into three parts:

1. Fine-tune the pre-trained vision-language model on data of the down-stream task **in English**

2. Fine-tune the model on data in the **auxiliary language** (one language other than English) using our proposed XVL-MAML algorithm.

3. Evaluate the fine-tuned model on data in the **target languages** (languages other than English and the auxiliary language).

The traditional cross-lingual transfer learning procedure described in Bugliarello et al. (2022) only includes part 1 and 3. In part 3, if the setting is zero-shot, the model is evaluated on data in the target language directly, but if the setting is few-shot, the model continues to be fine-tuned on few-shot data in the target languages and is then evaluated. The difference between our framework and the traditional procedure is the additional fine-tuning step of part 2. We will describe it specifically in Section 3.4, but before that, we will introduce contrastive learning for vision-and-language tasks.

3.3 Contrastive Learning for Vision-and-language Tasks

The vision-and-language contrastive learning loss proposed by Zhang et al. (2020) has proven effective in medical image scenarios and is used as the pre-training objective function of CLIP (Radford et al., 2021). It can be regarded as an auxiliary task for representation learning, aiming to enable models to gain better aligned multi-modal representation for downstream tasks. In the contrastive learning scheme, a batch of embeddings of images encoded by the model can be written as $I = \{I_1, ..., I_N\}$, and a batch of embeddings of texts encoded by the model can be written as $T = \{T_1, ..., T_N\}$, where N is the size of batch, (I_i, T_i) is an image-text pair. If the paired image-text data describe the same or similar concepts, then we can assume they are **positive** examples, and non-paired data are **negative** examples. Then, the embeddings of images and texts are fed into two different linear transformation layers separately, W_1 and W_2:

$$U = I \cdot W_1^\top \qquad (1)$$

$$V = T \cdot W_2^\top \qquad (2)$$

Where U and V represent the batch of image-text pairs. Then the cosine similarity of each pair can be computed as $\langle U_i, V_j \rangle = \frac{U_i^\top V_j}{\|U_i\|\|V_j\|}$. The objective is to maximize the similarity of matched image-text pairs and minimize the similarity of others. So the image-text contrastive loss can be formulated as follows:

$$\mathcal{L}_i^1 = -\log \frac{\exp(\langle U_i, V_i \rangle)}{\sum_{K=1}^N \exp(\langle U_i, V_k \rangle)} \qquad (3)$$

Following Zhang et al. (2020), the contrastive loss should be symmetric for each modality, and the text-image contrastive loss is:

$$\mathcal{L}_i^2 = -\log \frac{\exp(\langle V_i, U_i \rangle)}{\sum_{K=1}^{N} \exp(\langle V_i, U_k \rangle)} \quad (4)$$

The final contrastive loss of this batch of paired data is then:

$$\mathcal{L}_{CL} = \sum_{i=1}^{N} (\mathcal{L}_i^1 + \mathcal{L}_i^2) \quad (5)$$

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

3.4 XVL-MAML

Inspired by the effectiveness of MAML for quickly adapting to new tasks, we propose a novel variant of the MAML algorithm specialized for cross-lingual transfer in vision and language tasks, called XVL-MAML. Specifically, we first integrate contrastive learning into the MAML algorithm, making it specialized for the visual-language task of cross-lingual transfer learning. Our intuition is that we can use MAML with a contrastive loss as its learning objective for quickly adapting vision-language alignment to new languages. In this framework, the alignment between image and text in a specific language can be regarded as a task. Inspired by Nooralahzadeh et al. (2020), we use the data of one auxiliary language for fine-tuning, but with a contrastive loss as objective function in the MAML algorithm.

Specifically, we sample a batch of support data $\mathcal{B}_s$ and a batch of query data $\mathcal{B}_q$ in the data in auxiliary language A for each virtual task $\mathcal{T}$. Assuming the parameters of the model are θ and the contrastive loss on the support data is $\mathcal{L}_{CL}(\theta)_{\mathcal{B}_s}$, then the parameters of the model can be updated by one step of gradient descent:

$$\theta' = \theta - \alpha \nabla_\theta \mathcal{L}_{CL}(\theta)_{\mathcal{B}_s} \quad (6)$$

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

$$\theta \leftarrow \theta - \beta \nabla_\theta \mathcal{L}_{CL}(\theta')_{\mathcal{B}_q} \quad (7)$$

$$\theta \leftarrow \theta - \beta \nabla_\theta \mathcal{L}_{CL}(\theta - \alpha \nabla_\theta \mathcal{L}_{CL}(\theta)_{\mathcal{B}_s})_{\mathcal{B}_q} \quad (8)$$

Optimized using this method, pre-trained vision-language models can quickly adapt to new tasks

in other languages without using any annotation in the auxiliary language for downstream tasks, so we will refer to this as an **unsupervised scenario**.

In **supervised scenarios**, where the downstream tasks labels in the auxiliary language are available, we combine the loss of the downstream task $\mathcal{L}$ with the vision-language contrastive loss $\mathcal{L}_{CL}$ by adding them together. So during fine-tuning, Equation (8) is modified to:

$$\theta \leftarrow \theta - \beta(\nabla_\theta \mathcal{L}(\theta'')_{\mathcal{B}_q} + \lambda \nabla_\theta \mathcal{L}_{CL}(\theta')_{\mathcal{B}_q}) \quad (9)$$

Where the temporary parameters optimized for one step by the downstream task loss $\mathcal{L}$ on the support set $\mathcal{B}_s$ is θ'', β is the meta-learning rate, and λ is the scale factor of contrastive learning. By simply adding the gradients of the downstream task and contrastive learning in the meta-update, the model learns downstream tasks and vision-language alignment simultaneously for cross-lingual transfer.

4 Experiments

In this section, we introduce both the base PVLMs we use for vision-language cross-lingual transfer, as well as the datasets and metrics we use to evaluate our proposed method. Then we describe how the experiments were conducted and discuss the results.

4.1 Base models

In this paper, we choose xUNITER (Liu et al., 2021) and UC2 (Zhou et al., 2021) as our base models, as they use different pre-training methods. Then we applied XVL-MAML to both models to show that this method works across different models.

xUNITER is a multilingual version of the UNITER model (Chen et al., 2020). It has a similar architecture to UNITER and uses Faster-RCNN (Ren et al., 2015) as a feature extractor for images. The image features are pooled and reshaped as vectors with the same dimensions as text embeddings. UNITER has four pre-training methods: Masked Language Modelling (MLM), Masked Region Modelling (MRM), Image-Text Matching (ITM), and Word Region Alignment (WRA). xUNITER, in addition to these pre-training methods, also uses Masked Language Modelling for multilingual data and uses the same text embedder as XLM-R (Conneau et al., 2020).

| Method | Model | XNVLI | xGQA | MaRVL | xFlickr&Co | |
					IR	TR
Baseline	mUNITER	53.7	10.0	53.7	8.1	8.9
	xUNITER	59.0	20.8	56.0	13.8	12.5
	UC2	62.5	29.0	56.4	19.7	17.0
	M3P	58.2	28.2	56.0	12.9	11.9
Ours	xUNITER	63.0 (**+4.0**)	22.5 (**+1.7**)	**59.4** (**+4.4**)	16.3 (**+2.5**)	14.2 (**+1.7**)
	UC2	**64.4** (+1.9)	**29.9** (+0.9)	57.0 (+0.6)	**21.3** (+1.6)	**18.7** (+1.7)

Table 1: Zero-shot performance (accuracy) of four baseline models only fine-tuned on English data (Baseline) and two models fine-tuned by our meta-learning method (Ours) on four IGLUE datasets (Bugliarello et al., 2022).

UC2 uses a similar model architecture as UNITER, but different pre-training methods. Specifically, UC2 augments pre-training on English data by constructing a multilingual corpus via machine translation and then uses this augmented data for pre-training. It also proposes the 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.

4.2 Datasets and Metrics

We use datasets for four tasks from the IGLUE benchmark (Bugliarello et al., 2022), which includes xGQA (Pfeiffer et al., 2022), MaRVL (Liu et al., 2021), XVNLI, and xFlickr&Co (Plummer et al., 2015; Lin et al., 2014). We show examples from MaRVL and XVNLI in Figure 1. Following the convention in IGLUE, the evaluation metric is accuracy for all tasks except cross-modal retrieval, which uses Recall@1. The task format of these four datasets are described below:

- **MaRVL** is a multicutural vision-language reasoning dataset, following the format of English NLVR2 (Suhr et al., 2019) which namely to judge whether a sentence is correct or not for a pair of images.

- **XVNLI** is a multilingual version of visual natural language inference task, which requires models to predict the relationships between premise and hypothesis based on a given image.

- **xGQA** is a multilingual grounded question answering task based on GQA (Hudson and Manning, 2019) and machine translated question-answer pairs.

- **xFlickr&CO** is a multilingual image-text retrieval dataset collected from Flickr30k (Plum-

mer et al., 2015) and COCO (Lin et al., 2015)

4.3 Implementation and Hyperparameters

We conduct all experiments based on the Visiolinguistic Transformer Architectures framework VOLTA on four 2080Ti GPUs. We implement the MAML algorithm using the Higher library. We use the AdamW (Loshchilov and Hutter, 2018) optimizer to fine-tune all models in PyTorch.

Fine-tuning on English Data Before evaluating models on data in low-resource languages, we firstly fine-tune the pre-trained models on the corresponding English datasets: GQA (Hudson and Manning, 2019), NLVR2 (Suhr et al., 2019), SNLI-VE (Xie et al., 2019), and Flickr30k (Plummer et al., 2015) for xGQA, MaRVL, XVNLI, and xFlickr&Co, respectively, using the procedure of Bugliarello et al. (2022) and Liu et al. (2021). We follow the setting in IGLUE (Bugliarello et al., 2022) and also use the IGLUE hyper-parameters for each task when fine-tuning. We save the parameters of models in each epoch, then pick the best performing model for each task as the initialized parameters θ for the meta-learning fine-tuning stage.

Fine-tuning with Meta-learning For the XVL-MAML algorithm, the size of the support set and the query set is 64. We explore learning rates $5 \times 10^{-5}, 1 \times 10^{-5}, 5 \times 10^{-6}, 1 \times 10^{-6}$ for both UC2 and xUNITER, and find the best learning rate is 5×10^{-6} for both the normal fine-tuning stage and the meta-update of MAML. For the inner learning rate of XVL-MAML, we explore learning rates $5 \times 10^{-6}, 5 \times 10^{-5}, 5 \times 10^{-4}$ and 5×10^{-3}, and find that 5×10^{-4} is the best inner learning rate.

For the proposed meta-learning framework, we find that models overfit after 300 iterations in most situations (for each iterations, we sample a batch of data as support set and a batch as query set),

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 2: Zero-shot performance (accuracy/consistency) of two baseline models fine-tuned only on English data (Base) and then fine-tuned by our meta-learning method (Ours) on the MaRVL dataset (Liu et al., 2021), where the definition of consistency following Liu et al. (2021). Columns indicate target languages. The avg column gives the average performance across all target languages in this row. $zh \rightarrow X$ means the auxiliary language is Chinese, and the target languages is other low-resource languages X. We also show the average and maximum performance across all auxiliary languages for each target language.

so we set the number of iterations to 400 for all our 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 Results and Discussion

5.1 Zero-shot

We report the results of the baseline models and the results for fine-tuning them using our meta-learning framework in Table 1. In our setting, baseline model means that the PVLM is only fine-tuned on the English datasets. For simplicity, we report the averaged results of all combinations of target languages and auxiliary languages for each model and task. We set the value of λ in Equation (8) to 2×10^{-2} for xUNITER and 5×10^{-2} for UC2 to gain the best performance.

The results in the Table 1 indicate the effectiveness of our meta-learning framework and show that our method can boost the zero-shot performance of UC2 and XUNITER on all four datasets in IGLUE. Note that Table 1 shows average performance across all languages. The performance for individual languages can vary, and is shown in detail in Appendix A, Table 4. We also show the differences in improvements when using different auxiliary languages for different target languages in Figure 5.

5.2 Few-shot

We also conduct few-shot experiments following the setting in IGLUE (Bugliarello et al., 2022) for

Unsupervised Setting		
Method/Models	UC2	xUNITER
Baseline	$62.5_{\pm 0.1}$	$59.1_{\pm 0.1}$
XVL-MAML(w/o down-stream)	$\mathbf{63.1_{\pm 0.1}}$	$\mathbf{60.8_{\pm 0.1}}$
Supervised Setting		
Method/Models	UC2	xUNITER
XVL-MAML(w/o contrastive)	$63.8_{\pm 0.1}$	$61.6_{\pm 0.1}$
XVL-MAML	$\mathbf{64.4_{\pm 0.1}}$	$\mathbf{62.9_{\pm 0.1}}$

Table 3: Ablation study in the unsupervised setting and supervised setting. The labels of the down-stream task data in the auxiliary language are not given in unsupervised setting and provided in supervised setting.

both xUNITER and UC2 on XVNLI and MaRVL. The results are shown in Figure 2, where the horizontal axis represents the number of shots, and the vertical axis represents the accuracy score. The leftmost point of the horizontal axis is zero, which represents the performance in the zero-shot setup. The blue points and lines show the performance of our method. The yellow points and lines represent the performance of the baseline. We have performed five runs and the interval represents the standard error. It is clear that in all four figures, our method achieves better performance across 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, without seeing any data in the target languages, outperforms the baselines in the few-shot setting, except for UC2 on MaRVL. In other words, only a few instances of training data in target languages are not enough to eliminate the advantage of our method. This
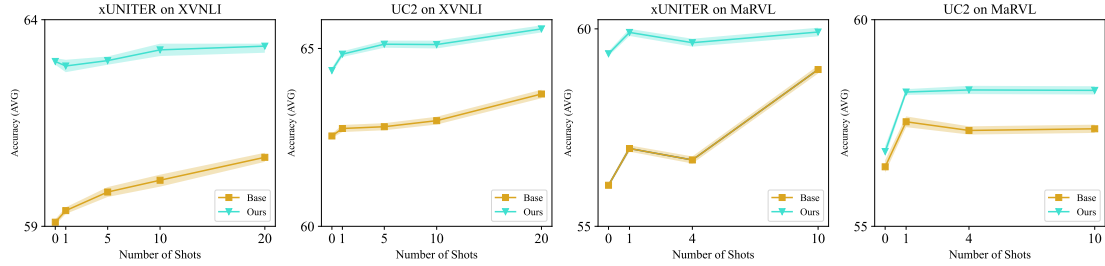
Figure 2: Average few-shot performance (accuracy) across all languages of two baseline models on the XVNLI and MaRVL datasets. The horizontal axis represents the number of shots in the training data.
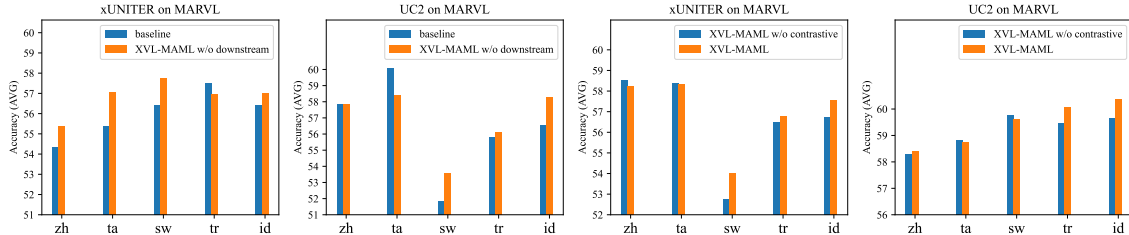

Figure 3: Performance in each target languages averaged across auxiliary languages on the MaRVL dataset.

demonstrates that while our method requires training data in one auxiliary language, there is no need for few-shot data in the target languages.

5.3 Ablation Study and Further Analysis

In this section, we conduct a series of ablation studies which investigate the effect of each part of our proposed meta-learning framework. We have performed five runs for each setting and reported the average and standard error to estimate significant differences.

The Effect of Contrastive Learning We investigate the effect of contrastive learning in our meta-learning fine-tuning framework. Specifically, we fine-tune the model only using a contrastive learning loss in the MAML algorithm (called as "XVL-MAML (w/o down-stream)" in Table 3), where the labels of down-stream task data are not given. We evaluate the performance of UC2 and xUNITER on the XVNLI dataset in this setting and reported them in unsupervised setting part of Table 3. The results indicate that using contrastive learning solely in the MAML algorithm can improve performance. It provides evidence for the hypothesis that contrastive learning can enable models to learn alignments of modalities in cross-lingual transfer, resulting in better representations.

We also compare the performance of the model in the supervised setting where labels of data in auxiliary language are available; hence in the XVL-MAML algorithm, both contrastive loss and down-stream task loss are used. Then we remove the contrastive learning loss in XVL-MAML, only keeping the down-stream task loss. We compare the performance of these two settings in Table 3 to show the effectiveness of the contrastive learning loss in XVL-MAML in the supervised setting. In the "Supervised Setting" part of Table 3, the first row is XVL-MAML without contrastive learning loss, which means only using down-stream task loss when fine-tuning, and the second row is normal XVL-MAML using both contrastive loss and down-stream task loss.

Moreover, we show the difference in performance in each target language separately in Figure 3. Contrastive learning can bring improvements for most of the target languages, especially those whose performance is relatively low when not using contrastive learning. For example, in the left-most plot, performance in *zh*, *ta*, and *sw* is relatively lower than *tr* in the baseline, but gains significant improvements when using our method. The similar effect can be seen in other three plots and Table 2.

Diverse down-stream tasks We report the results of experiments using four diverse multilingual vision-and-language understanding tasks in Table 1. Our method can bring clear improvements across all tasks for both UC2 and xUNITER, indicating that the approach generalises across tasks. Furthermore, these four IGLUE tasks also differ in the distribution of language families and domains, which indicates our method can be useful across
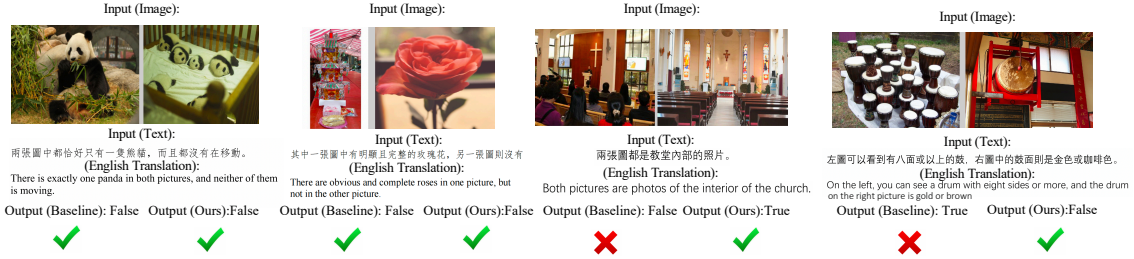
18

Figure 4: Examples from the Chinese part of the MaRVL dataset and predictions of the baseline and ours method.
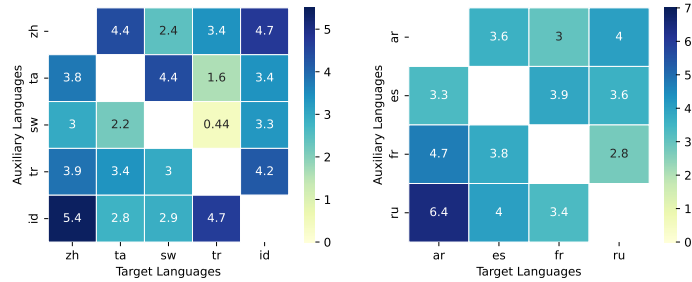

Figure 5: Improvements of zero-shot performance by fine-tuning xUNITER on different auxiliary languages then evaluating on different target languages using our proposed framework compared with baseline. The left heatmap is for MaRVL, and the right is for XVNLI. Rows correspond to auxiliary and columns correspond to target languages.

language families and domains. Moreover, our method can significantly boost the performance of xUNITER even in the challenging MaRVL dataset which encompasses five diverse language families and cultures, improving accuracy by 4.4 points.

Diverse languages We also investigate the difference of performance between languages. Specifically, we take the MaRVL dataset as an example and report results in Table 2, which lists the performance when using Chinese (zh) as the auxiliary language for meta-learning, and the average and maximum performance across all auxiliary languages for each target language respectively. In most situations, our method results in clear improvements. We then visualize the improvements of xUNITER when using different auxiliary languages for different target languages on MaRVL and XVNLI in Figure 5. The improvements we see for MaRVL (which range from 0.44 to 5.4) are smaller than for XVNLI (which range from 2.8 to 6.4), and one possible reason is that the language families of MaRVL are more diverse than those of XVNLI. But in general, our method improves performance for all combinations of auxiliary and target languages, even when they come from different language families. This further indicates that our method is language-agnostic.

5.4 Example Predictions

We show some examples of inputs and predictions for baseline and our method in Figure 4. We use xUNITER to predict the Chinese part of the MaRVL dataset. We have selected two examples where the baseline prediction is incorrect, but our method predicts correctly (the rightmost two examples), and two examples where both our method and baseline method predict correctly (the leftmost two examples). In the two rightmost examples, the label is "True", but the baseline predicts "False". We find that in these two examples, the same concepts ("church" and "drum") described in related texts have different visual features, which makes it more difficult for models to identify them. In the left two examples, however, the concepts (panda and roses) described in the text do not have diverse or obscure visual features when they appear in the images. Therefore, based on these cases, we can surmise that the meta-learning framework makes the model more adaptive to diverse information and resulting in better generalization capabilities when mapping between texts and images.

6 Conclusions

In this paper, we focused on mitigating the problem of poor performance of current PVLMs in vision-language cross-lingual transfer. We proposed a novel MAML framework to adapt pre-trained mod-

els for new languages in vision-and-language tasks. Our framework combines contrastive learning and downstream task supervised learning. We verify the effectiveness of our approach in both supervised and unsupervised settings. The key strength of our method is that we leverage contrastive learning in the MAML procedure so that models can quickly learn to align representations from different modalities and adapt them to unseen languages.

Experimental results demonstrate that our proposed meta-learning framework significantly improves the performance of models in vision-and-language cross-lingual transfer both in zero-shot and few-shot setups. We applied our method to two representative PVLMs, UC2 and xUNITER, and verified its effectiveness on four datasets in the IGLUE benchmark in 14 languages. We also conducted an ablation study to explore the effect of contrastive learning, and analysed the effect of different languages and tasks.

Limitations

Our proposed method applies contrastive learning to samples of image-text pairs. The alignments induced in this fashion work best if there is a concept or an object that is both depicted in the image and referred to in the sentence. If this is not the case, then the method may end up learning incorrect alignments; this includes cases where the image or the sentence contain multiple objects or concepts, not all of which can be aligned. To address this limitation, future work should explore how to construct better positive and negative samples and how to enable learning at a more fine-grained level. Besides, current famous PVLMs are encoder-only models, which is different with recent decoder-only LLMs, so meta-learning methods for multi-modal multilingual LLMs is worth to explore as a future work.

Ethics Statement

The use of the IGLUE benchmark in our paper is consistent with its intended use. We have checked the datasets for offensive content by sampling and visualizing examples. There are 14 languages in the datasets we use, we list them in Table 4. More detailed information about the IGLUE dataset can be found in (Bugliarello et al., 2022).

References

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

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. Uniter: Universal image-text representation learning. In *European conference on computer vision*, pages 104–120. Springer.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Édouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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.

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

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

Spandana Gella, Rico Sennrich, Frank Keller, and Mirella Lapata. 2017. Image pivoting for learning multilingual multimodal representations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2839–2845, Copenhagen, Denmark. Association for Computational Linguistics.

Jiatao Gu, Yong Wang, Yun Chen, Victor O. K. Li, and Kyunghyun Cho. 2018. Meta-learning for low-resource neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3622–3631, Brussels, Belgium. Association for Computational Linguistics.

Drew A Hudson and Christopher D Manning. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709.

Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy Hospedales. 2018. Learning to generalize: Meta-learning for domain generalization. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. 2015. Microsoft coco: Common objects in context.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer.

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

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

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32.

Minheng Ni, Haoyang Huang, Lin Su, Edward Cui, Taroon Bharti, Lijuan Wang, Dongdong Zhang, and Nan Duan. 2021. 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.

Farhad Nooralahzadeh, Giannis Bekoulis, Johannes Bjerva, and Isabelle Augenstein. 2020. 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.

Abiola Obamuyide and Andreas Vlachos. 2019. Model-agnostic meta-learning for relation classification with limited supervision. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5873–5879.

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

Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. 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.

Chen Qiu, Dan Oneata, Emanuele Bugliarello, Stella Frank, and Desmond Elliott. 2022. Multilingual multimodal learning with machine translated text. *arXiv preprint arXiv:2210.13134*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR.

Mengye Ren, Eleni Triantafillou, Sachin Ravi, Jake Snell, Kevin Swersky, Joshua B Tenenbaum, Hugo Larochelle, and Richard S Zemel. 2018. Meta-learning for semi-supervised few-shot classification. *arXiv preprint arXiv:1803.00676*.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28.

Guy Rotman, Ivan Vulić, and Roi Reichart. 2018. Bridging languages through images with deep partial canonical correlation analysis. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 910–921, Melbourne, Australia. Association for Computational Linguistics.

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

Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. 2019. 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.

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

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

Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. 2016. Matching networks for one shot learning. *Advances in neural information processing systems*, 29.

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

Bailin Wang, Mirella Lapata, and Ivan Titov. 2021. Meta-learning for domain generalization in semantic parsing. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 366–379.

Ying Wang, Jonas Pfeiffer, Nicolas Carion, Yann LeCun, and Aishwarya Kamath. 2023. Adapting grounded visual question answering models to low resource languages. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2595–2604.

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

Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. 2020. Gradient surgery for multi-task learning. *Advances in Neural Information Processing Systems*, 33:5824–5836.

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

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

Mingyang Zhou, Luowei Zhou, Shuohang Wang, Yu Cheng, Linjie Li, Zhou Yu, and Jingjing Liu. 2021. 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.

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

A   Appendix

This is a section in the appendix.

	Ar	Bn	De	Es	Id	Fr	Ja	Ko	Pt	Ru	Sw	Ta	Tr	Zh
					MaRVL									
xUNITER (Baseline)	-	-	-	-	56.44	-	-	-	-	-	56.41	55.40	57.53	54.34
UC2 (Baseline)	-	-	-	-	56.56	-	-	-	-	-	51.81	**60.06**	55.76	57.81
xUNITER (Ours)	-	-	-	-	**60.35**	-	-	-	-	-	**59.59**	58.49	**60.06**	59.75
UC2 (Ours)	-	-	-	-	56.74	-	-	-	-	-	54.60	58.94	58.13	59.59
					XVNLI									
xUNITER (Baseline)	53.52	-	-	60.05	-	61.60	-	-	-	61.25	-	-	-	-
UC2 (Baseline)	58.36	-	-	**63.86**	-	65.01	-	-	-	64.72	-	-	-	-
xUNITER (Ours)	56.70	-	-	60.91	-	68.64	-	-	-	63.91	-	-	-	-
UC2 (Ours)	**59.94**	-	-	62.97	-	**69.41**	-	-	-	**65,18**	-	-	-	-
					xGQA									
xUNITER (Baseline)	-	11.41	33.21	-	32.38	-	-	13.28	20.51	17.84	-	-	-	17.20
UC2 (Baseline)	-	19.49	33.52	-	29.83	-	-	23.29	31.23	32.61	-	-	-	33.25
xUNITER (Ours)	-	12.46	34.10	-	**33.63**	-	-	15.05	22.71	20.27	-	-	-	19.27
UC2 (Ours)	-	**19.63**	**34.50**	-	29.58	-	-	**24.93**	32.47	33.24	-	-	-	**35.00**
					Xflickr&Co (IR)									
xUNITER (Baseline)	-	-	14.70	16.40	15.15	-	9.55	-	-	14.75	-	-	8.85	17.20
UC2 (Baseline)	-	-	28.10	14.65	13.55	-	23.70	-	-	18.20	-	-	8.15	31.70
xUNITER (Ours)	-	-	16.20	**18.85**	18.50	-	12.10	-	-	17.75	-	-	**11.10**	19.40
UC2 (Ours)	-	-	**29.35**	16.90	14.25	-	**25.15**	-	-	**20.50**	-	-	10.50	**32.10**
					Xflickr&Co (TR)									
xUNITER (Baseline)	-	-	14.2	15.45	13.95	-	8.30	-	-	13.15	-	-	7.75	14.4
UC2 (Baseline)	-	-	23.55	11.90	10.35	-	22.75	-	-	17.50	-	-	6.15	26.85
xUNITER (Ours)	-	-	15.50	**16.15**	**16.70**	-	9.90	-	-	15.70	-	-	**9.50**	15.75
UC2 (Ours)	-	-	**25.30**	13.95	12.45	-	**23.50**	-	-	**19.80**	-	-	8.30	**27.45**

Table 4: Accuracy scores for each target language individually averaged over auxiliary languages.