Parameter-Efficient Cross-Lingual Transfer of Vision and Language Models via Translation-based Alignment

Anonymous ACL submission

Abstract

Pre-trained vision and language models such as CLIP (Radford et al., 2021) have witnessed remarkable success in connecting images and texts with a primary focus on English texts. Despite recent efforts to extend CLIP to support other languages, disparities in performance among different languages have been observed due to uneven resource availability. Additionally, current cross-lingual transfer methods of those pre-trained models would consume excessive resources for a large number of languages. Therefore, we propose a new parameter-efficient cross-lingual transfer learning framework that utilizes a translation-based alignment method to mitigate multilingual disparities and explores parameter-efficient fine-tuning methods for parameter-efficient cross-lingual transfer. Extensive experiments on XTD (Aggarwal and Kale, 2020) and Multi30K (Elliott et al., 2016) datasets, covering 11 languages under zero-shot, few-shot, and full-dataset learning scenarios, show that our framework significantly reduces the multilingual disparities among languages and improves cross-lingual transfer results, especially in low-resource scenarios, while only keeping and fine-tuning an extremely small number of parameters compared to the full model (e.g., Our framework only requires 0.16% additional parameters of a full-model for each language in the few-shot learning scenario).

1 Introduction

Cross-lingual transfer is the ability of a model to utilize knowledge from high-resource language to improve its performance in low-resource language. This is particularly useful in situations where data is limited or costly to collect in certain languages. Cross-lingual transfer has been applied to various NLP tasks such as sentiment classification (Chen et al., 2018), dependency parsing (Ahmad et al., 2018), named entity recognition (Rahimi et al., 2019), question answering (Lewis et al., 2019), and dialog (Schuster et al., 2018). Recent works, including XLM-R (Conneau et al., 2019), mBART (Liu et al., 2020), and mT5 (Xue et al., 2020), have extended language models to a multilingual version using this technique.

Two-stream vision-language pre-trained model CLIP (Radford et al., 2021) has demonstrated remarkable performance in image-text retrieval (Cao et al., 2022) by encoding images and text into a shared representation space. However, it primarily focuses on English and cannot comprehend other languages. To address this limitation, Multilingual-CLIP (Carlsson et al., 2022a) has been proposed to enhance the CLIP's ability to support multiple languages through cross-lingual transfer. Nevertheless, Multilingual-CLIP treats English as a pivot language, leading to performance disparities across languages, especially low-resource languages. While previous work (Wang et al., 2022) has accessed and highlighted this multilingual disparity, there is currently a lack of proposed solutions to address it.

Moreover, multilingual models often face a trade-off across different languages (Xin et al., 2022), meaning that the model's capabilities in one language may degrade its performance in another. This can be a significant issue as the need to train and maintain separate models for each language can become resource-intensive when dealing with a large number of languages.

In order to better understand and solve the above problems, we are interested in the following research questions.

RQ1: Is cross-lingual transfer learning necessary for CLIP? The original CLIP model with machine translation can also solve the multilingual problem to a certain extent. Comparing the original and multilingual CLIP models can reveal the current multilingual capabilities and disparities of the multilingual CLIP and guide efforts to reduce multilingual disparities. *RQ2: How to exploit pivot lan-*

guage for better cross- lingual transfer and reduce the multilingual disparity? Compared with low-resource languages, texts in pivot language (English) are much easier to be obtained and cheaper to be annotated (Pavlick et al., 2014). Therefore, it is easy to get pivot-target text pair describing the same image in different languages. Since the model has better performance on pivot language and pivot-target text pairs can provide more information, there must be a better approach to exploit pivot language for better cross-lingual transfer of Multilingual-CLIP and to help reduce the disparity. *RQ3: Can we do the cross-lingual transfer of Multilingual-CLIP in a parameter-efficient manner?* Saving and deploying the entire model for each language is very resource-consuming. Therefore, we need a parameter-efficient solution that uses fewer additional parameters and does not result in large performance drops.

In this paper, we aim to reduce the multilingual disparity in a parameter-efficient way. To achieve this, we propose a framework building upon the Multilingual-CLIP model. To be specific, in the framework, we propose a translation-based alignment method to narrow the distribution gap between translation and natural language distribution, which significantly reduces the multilingual disparity of Multilingual-CLIP. Additionally, we adopt Parameter-Efficient Tuning (PET) methods (Houlsby et al., 2019; Karimi Mahabadi et al., 2021; He et al., 2022a; Rücklé et al., 2020; Li and Liang, 2021; Guo et al., 2020; Hu et al., 2021; Zaken et al., 2021; Lester et al., 2021a) as a solution to achieve parameter efficiency. Furthermore, we find that, in the zero-shot scenario, hard prompt can also reduce the multilingual disparity and improve multilingual ability in addition to parameter efficiency. Compared with full-model fine-tuning on each language, our framework mitigates the multilingual disparity and obtains higher average performance across all languages, using much fewer additional parameters than a single model.

We conduct our experiments on XTD and Multi30K datasets covering 11 languages in zero-shot, few-shot, and full-dataset learning scenarios. Through extensive analytical experiments, we verify the effectiveness of our framework and provide answers to our research questions. From our experimental results, conclude the following:

1. The Multilingual-CLIP model can achieve better performance than the original CLIP model, but still suffers from a significant multilingual disparity. Meanwhile, we find mapping the text embedding to a better distribution using machine translation can reduce this multilingual disparity. (Section 5.2)

2. Mapping the text embedding to a better distribution and approximating it to natural pivot language distribution can significantly help reduce the multilingual disparity. (Section 5.3)

3. PET methods can help address the excessive resource consumption of Multilingual-CLIP and the performance degradation is within an acceptable range. PET methods can address the excessive resource consumption of Multilingual-CLIP while maintaining acceptable performance degradation. Moreover, we find that hard prompt in English is very effective in the zero-shot learning scenario and can be applied to all languages. (Section 5.4)

2 Background

2.1 Multilingual-CLIP

CLIP (Radford et al., 2021), proposed by OpenAI, is a two-stream vision-language pre-trained model with textual and visual encoder. It is trained on a large scale image-text pair dataset using a contrastive loss to encode the image and text into a shared embedding space. CLIP calculate the cosine similarity between image and text features to measure their semantic similarity.

Recently, Multilingual-CLIP (Carlsson et al., 2022b) extend CLIP to a multilingual version. This work replaces original English text encoder with a pre-trained multilingual language model such as M-BERT (Devlin et al., 2018) and trains it using teacher learning (Hinton et al., 2015). Although Multilingual-CLIP endowed CLIP with multilingual capabilities, the performance of Multilingual-CLIP in other languages is worse than in English due to the limited amount of data available in low-resource languages, leading to insufficient training in these languages. Furthermore, training data for other languages are translated from English text, which can result in a distribution gap during training and practical application. Noticing this problem, we aim to reduce this multilingual disparity in this paper.

2.2 Parameter-Efficient Tuning

As the size of foundation models (Bommasani et al., 2021) increases, fine-tuning and saving the entire model becomes very resource-intensive. Many parameter-efficient tuning (PET) methods have been proposed to solve this issue. These approaches add additional parameters inside the model (Houlsby et al., 2019; Karimi Mahabadi et al., 2021; He et al., 2022a; Rücklé et al., 2020; Li and Liang, 2021), optimize a small portion of the parameters or their low-rank decomposition matrix(Guo et al., 2020; Hu et al., 2021; Zaken et al., 2021), or add trainable token embedding into the input (Lester et al., 2021a). Moreover, He et al. (2021) and Ding et al. (2022) analyze and combine these approaches from a unified perspective. Furthermore, Hu et al. (2022) and Zhang et al. (2022) propose automatic methods to search for an optimal combination of these pet methods for language models and visual models, respectively. Many works (Gao et al., 2021; Zhou et al., 2022; Zhang et al., 2021; He et al., 2022b) also apply PET methods to CLIP models. Nevertheless, those PET methods have not been thoroughly explored for the Multilingual-CLIP model in the cross-lingual transfer setting. It is important to note that PET methods often result in a decline in performance to varying degrees compared to full-model fine-tuning. Therefore, it is essential to conduct experiments to verify the effectiveness of these methods and determine the most appropriate approach for specific tasks and models.

3 Framework

Our main contribution is proposing a cross-lingual transfer framework for Multilingual-CLIP (Figure 1). In this framework, we propose a novel translation-based cross-lingual alignment method to reduce the multilingual disparity and exploit parameter-efficient tuning methods to solve the resource consumption problem in cross-lingual transfer.

3.1 Translation-based Cross-lingual Alignment

In Figure 1(b), we present a diagram of translation-based cross-lingual alignment method. Blue circles represent the distribution of natural language embeddings in the representation space, while orange represents the embedding distribution of text generated by machine translation. Since Multilingual-CLIP trains in other languages in the pre-training process by only aligning the English text with the target text translated from English, there is a gap in text distribution between training process and practical applications. The gap for the various languages is different, which contributes to the multilingual disparity. In our framework, we use machine translation to map one embedding distribution to another and propose an alignment method on pivot-target language text pair, which describe the same image in pivot (English) and target language, to narrow the distribution gap. Our alignment method has different combinations of routines and loss functions to be compared.

Alignment Routines. Given the pivot language (English) and target language, there are three routines in the representation space to narrow the gap between embedding distributions as shown in the left part of Figure 1. To be specific, these routines are (1) aligning original English and target language text embeddings, (2) aligning translated English (to target) and original target language text embeddings, as Multilingual-CLIP is only pre-trained in target language in translation distribution where it performs better than nature language distribution, and (3) aligning original English and translated target language (to English) text embeddings. We compare these three routines in the experiments and find routine 3 performs best. Note that these three routines do not apply to pivot languages (English) and routine 3 still need machine translation in the inference process.

Alignment loss functions. In addition to the alignment routines, alignment loss functions must also be considered. Mean Squared Error (MSE) loss and contrastive loss are two practical loss functions for narrowing the distance between embeddings.

To be specific, The original contrastive loss between image and text embeddings can be written as:

$$\mathcal{L}_{\text{i2t}} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\mathrm{e}^{\cos(v_i,t_i)/\tau}}{\sum_{j=1}^{N}\mathrm{e}^{\cos(v_i,t_j)/\tau}}, \quad (1)$$

$$\mathcal{L}_{\text{t2i}} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\mathrm{e}^{\cos(t_i,v_i)/\tau}}{\sum_{j=1}^{N}\mathrm{e}^{\cos(t_i,v_j)/\tau}}, \quad (2)$$

where v_i is the visual embedding of the image in the i-th pair and t_j represents the textual embedding of the text in the j-th pair. We use i2t and t2i

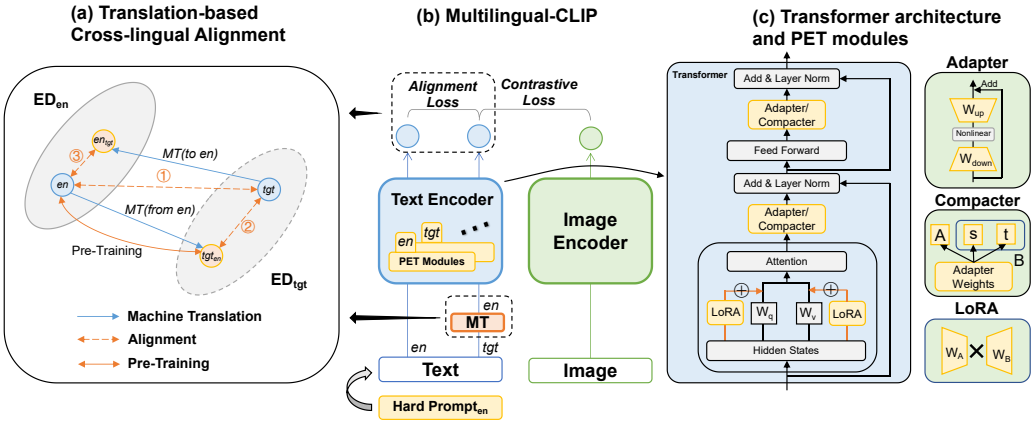

Figure 1: Illustration on our framework based on multilingual CLIP. we propose a translation-based alignment method to narrow the distribution gap and adopt PET methods to achieve parameter efficiency. We also find hard prompt is very effective in the zero-shot scenario.

to represent image-text and text-image matching. τ is the temperature used to scale the cosine similarity. Following CLIP (Radford et al., 2021), it is set to 0.01. N is the number of image-text pairs in the dataset. Pivot and target language text embedding can be obtained through text encoder:

$$T^{\text{pivot}} = \text{text\_encoder}(\text{text}^{\text{pivot}}), \quad (3)$$

$$T^{\text{tgt}} = \text{text\_encoder}(\text{text}^{\text{tgt}}), \quad (4)$$

Note that texts can be translated from one to another:

$$\text{text}^{\text{tgt}} \leftarrow \text{Trans}_{\text{tgt}\rightarrow\text{en}}(\text{text}^{\text{tgt}}), \quad (5)$$

$$\text{text}^{\text{pivot}} \leftarrow \text{Trans}_{\text{en}\rightarrow\text{tgt}}(\text{text}^{\text{pivot}}). \quad (6)$$

These two different losses to regularize the distance between parallel text embeddings can be represented as:

$$\mathcal{L}^{\text{MSE}}_{\text{alignment}} = \text{MSE}(T^{\text{pivot}}, T^{\text{tgt}}), \quad (7)$$

$$\mathcal{L}^{\text{contrastive}}_{\text{alignment}} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{\cos(t_i^{\text{pivot}}, t_i^{\text{tgt}})/\tau}}{\sum_{j=1}^{N} e^{\cos(t_i^{\text{pivot}}, t_i^{\text{tgt}})/\tau}}. \quad (8)$$

They are added to the contrastive loss between image and text with a alignment coefficient λ:

$$\mathcal{L} = \mathcal{L}_{\text{i2t}} + \mathcal{L}_{\text{t2i}} + \lambda \cdot \mathcal{L}_{\text{alignment}}. \quad (9)$$

3.2 Parameter-Efficient Cross-lingual Transfer Learning

We compare the following parameter-efficient tuning (PET) methods with full-model fine-tuning. When training these PET modules, we freeze the parameter of Multilingual-CLIP. PET methods are usually designed for different task. We instead use PET methods for different languages to achieve parameter efficiency.

Adapter (Houlsby et al., 2019): Figure 1(c) top right. Adapter adds a few trainable linear neural modules after every attention and feed-forward layer. it consists of a down sampling matrix $W_{down} \in \mathbb{R}^{d\times r}$ and a up sampling matrix $W_{up} \in \mathbb{R}^{r\times d}$ with a nonlinear activation function f in the middle, where d is the dimention of the input $x \in \mathbb{R}^d$ of the adapter. There is also a residual connection and the output O can be written as:

$$O = x + f(xW_{down})W_{up} \quad (10)$$

Compacter (Karimi Mahabadi et al., 2021): Figure 1(c) center right. An improvement work of Adapter. This work replaces the standard Adapter layer with a low-rank hypercomplex Adapter layer, which requires fewer parameters and yields competitive results. To be specific, Compacter decompose the $W_{down} \in \mathbb{R}^{d\times r}$ to the sum of k Kronecker products of matrix $A_i \in \mathbb{R}^{k\times k}$ and $B_i \in \mathbb{R}^{\frac{d}{k}\times\frac{r}{k}}$. B_i is further decomposed to two low-rank matrices $s_i \in \mathbb{R}^{\frac{d}{k}\times r_B}$ and $t_i \in \mathbb{R}^{r_B\times\frac{d}{k}}$, where r_B represents the rank of B_i. Finally, the formula can be written as follows:

$$W_{down} = \sum_i^k A_i \otimes B_i = \sum_i^k A_i \otimes (s_i t_i) \quad (11)$$

W_{up} is also decomposed in this way with shared A_i.

LoRA (Hu et al., 2021): Figure 1(c) bottom right. LoRA assumes the low-rank change of model weights $W \in \mathbb{R}^{d\times K}$ and then uses two trainable rank-decomposition matrices $W_A \in \mathbb{R}^{d\times r}$ and $W_B \in \mathbb{R}^{r\times k}$ to approximate the matrix change. Consequently, LoRA adds changes to the original output O from the input x:

$$O \leftarrow O + xW_AW_B \tag{12}$$

Following the default setting of LoRA, we apply this method to query and value projection matrices (W_q, W_v) in self-attention layers.

Hard and soft Prompt (Lester et al., 2021b): Figure 1(b) bottom. Hard prompt attaches text prompts to the front of the input text (e.g., "a photo of [Text]"), which is manually designed and explainable. CLIP uses multiple hard in the pre-training phase, so we are interested in whether it is applicable in cross-language transfer scenarios. We compare different combinations of the hard prompt and input text in different languages, and find English hard prompt work well on average across all languages. Soft prompt, also known as prompt tuning, adds trainable token embeddings to the front of the input. We do not plot soft prompt in our framework as our experiments show it doesn't perform well.

In our experiments, we also tune the linear head and layer-norm layer of the text encoder when we train Adapter, Compacter and LoRA and the number of their parameters is 0.44%, 0.05% and 0.16% of the text encoder, respectively. Hard prompts only require saving a few words, while soft prompts require storing several token embeddings, each of which takes up 1024 floating point numbers in storage space.

3.3 Optimization Objective

With the Translation-based alignment on pivot-target language text pairs and parameter-efficient tuning methods in cross-lingual scenario, our optimization objective is to get optimal parameters of PET modules on each target languages throught minimizing the loss $\mathcal{L}$. We denote the pivot-target language text pair training datasets as $\mathcal{D}^{pivot}$ and $\mathcal{D}^{target}$ and parameters of the PET modules for target language and freezed Multilingual-CLIP as θ^{PET}_{target} and θ^{MC}. Then our optimization objective can be formulated as follow:

$$\arg\min_{\theta^{PET}_{target}} \mathcal{L}(\mathcal{D}^{pivot}, \mathcal{D}^{target}; \theta^{PET}_{target}, \theta^{MC}). \tag{13}$$

4 Experimental Setup

Dataset We work with two datasets: (1) Flickr30K (Young et al., 2014) is an image captioning datasets in English, split into train/dev/test datasets with the number of 29000/1024/1000. The Multi30K (Elliott et al., 2016) dataset extends captions of Flickr30K dataset with human translated and independent German sentences. Elliott et al. (2017) and Barrault et al. (2018) further translate English Flickr30k captions to French and Czech, respectively. (2) XTD (Aggarwal and Kale, 2020) is a Cross-lingual dataset for the image-text retrieval task covering 11 languages. It only has a test split with 1000 samples per language with the same images. For few-shot setting, we randomly split the original test split into train/dev/test sets with 50/50/900 image-text pairs.

Base Model and Translation Tool We use XLM-R$_{Large}$-ViT$_{L/14}$ (Carlsson et al., 2022a) as our base model. The model fixes the original visual encoder of OpenAI ViT$_{L/14}$ (Radford et al., 2021) and replaces the text encoder by XLM-Roberta$_{Large}$ (Conneau et al., 2019) trained by teacher learning (Hinton et al., 2015). We do not evaluate MURAL (Carlsson et al., 2022b) as the code and model of MURAL are not available now. we use Google Translation[1], a strong Neural Machine Translation (NMT) system, to translate between all the different languages. To reduce the computational overhead, we translate the dataset in advance, rather than when it is used. We give the results of the translation in the code part.

5 Experiments and Analysis

In this section, we first present the overall results of our framework with the optimal combination and then conduct analytical experiments to demonstrate the effectiveness of machine translation (Section 5.2), routine 3 and MSE loss is the best choice for alignment (Section 5.3) and hard prompt, LoRA and Adapter is respectively outperforms in zero-shot, few-shot and full-dataset scenario (Secion 5.4). The details of our experimental configurations are in Appendix A.

5.1 Cross-Lingual Transfer Results on XTD and Multi30K

Table 1 shows the results of Multilingual-CLIP with and without our framework on XTD and Multi30K datasets in zero-shot, few-shot and full-dataset scenarios. Our framework adopts the opti-

[1]https://translate.google.com/

Framework	XTD				Multi30K			
	en↑	Avg.$_{\text{-en}}$ ↑	Std↓	Range↓	en↑	Avg.$_{\text{-en}}$ ↑	Std↓	Range↓
	Zero-shot							
w/o (M-CLIP)	63.44	57.42	4.75	16.00	66.65	65.03	0.93	2.15
w/ (Ours)	**64.06**	**59.94**	**3.26**	**10.84**	**67.80**	**66.73**	**0.59**	**1.30**
	Few-shot				Full-dataset			
w/o (M-CLIP)	64.67	58.57	4.83	16.06	**76.10**	74.93	0.70	1.55
w/ (Ours)	**64.83**	**60.45**	**3.10**	**10.61**	75.35	**75.65**	**0.56**	**1.25**

Table 1: Results on XTD and Multi30K of Multilingual-CLIP (M-CLIP) with and without our framework. We report the Recall@1 score and bold the best result in each scenario on each dataset. "Avg.$_{\text{-en}}$" represents the average score without English. Statistical indicators standard deviation (Std) and range are used to evaluate multilingual disparity.

mal combination of We report the Recall@1 score on English dataset and average score across other all languages to evaluate the multilingual performance and calculate statistical indicators, standard deviation and range, to evaluate the multilingual disparity. Compared with Multilingual-CLIP without our framework, our framework outperforms in all zero-shot, few-shot and full-dataset scenarios. It reduces range by more than 5 points and Std by more than 1.5 points while achieving significant performance improvement both in English and on average across all other languages on XTD dataset in both zero-shot and few-shot scenarios, which is a common application scenario of low-resource languages. The improvement on the Multi30K dataset is also significant. In terms of the number of parameters, our framework is also more efficient than full-model fine-tuning. In eleven languages, Adapter and LoRA only use 4.89% and 1.73% of the parameters respectively, which is far less than the parameters of 11 models.

In short, our framework significantly reduces the multilingual disparity and enables parameter-efficient cross-lingual transfer. We are also surprised to find that our framework improves multilingual performance.

5.2 Analysis on Multilingual Disparity of Multilingual CLIP

Since Multilingual-CLIP replaces the text encoder with a multilingual version while keeping the image encoder of CLIP, a direct evaluation is the comparison between CLIP and Multilingual-CLIP. Original CLIP with machine translation can be a strong baseline (Jain et al., 2021). We compare the performance of the CLIP and Multilingual CLIP models on XTD dataset with the help of machine translation. We first translate languages other than English into English by machine translation and then use it as input for CLIP and Multilingual-CLIP. For further validation, we also translated the English dataset to each language and tested them on multilingual CLIP.

Result analysis. Results are shown in Table 2. First, we can see that even though the original CLIP has limited multilingual capabilities, with the help of machine translation, it can achieve high multilingual capabilities to a certain extent and even surpass Multilingual-CLIP. However, Multilingual-CLIP with machine translation obtain the best multilingual capability and the lowest disparity in both scenarios. We did not use the setting "M-CLIP (en→tgt)" as a comparison as it is not a practical application scenario. Second, there is a large multilingual disparity for multilingual CLIP, e.g., the difference between Japanese and English is up to 16 and the standard deviation is up to 4.75. It indicates that the multilingual ability of multilingual CLIP is unbalanced, with a large room for improvement. With the help of machine translation, the improvement of Multilingual-CLIP on multilingual differences is very obvious, especially in languages that have performed poorly. Finally, using data translated from English, Multilingual-CLIP shows a large improvement in other languages (mean improvement of 1.6) but still lower than in English. This may be because the model is pre-trained with text translated from English, making it more adapted to this situation. This also suggests that multilingual disparity are partly the result of differences in the quality of the datasets in different language instead of the ability of model.

Explanation. From the perspective of representation space, machine translation maps text from one embedding distribution to another. Although the embedding distribution of the translated text is slightly different from that of the natural language, their difference is much smaller than that between the two different languages, which is also easier to be optimized to get closer.

5.3 Analysis on How to Exploit Pivot Language

Datasets in low-resource languages are usually small, and we can obtain the corresponding pivot language (English) text from target language text through human annotation. In particular, for the image caption dataset, annotators can directly give high-quality English captions based on the image

	Method	XTD														
		en	de	fr	es	it	ko	pl	ru	tr	zh	jp	Avg.↑	Avg.$_{-en}$↑	Std↓	Range↓
Zero-shot	CLIP	62.06	25.33	32.94	31.28	25.00	0.56	5.67	1.72	4.50	1.39	6.83	17.93	13.52	19.36	61.50
	CLIP + MT	62.06	60.56	60.56	59.72	58.78	58.00	60.11	53.72	60.06	56.72	50.72	58.27	57.90	3.38	11.34
	M-CLIP	**63.44**	59.94	60.06	58.90	60.72	51.00	**61.50**	56.11	59.28	59.28	47.44	57.97	57.42	4.75	16.00
	M-CLIP (en→tgt)	**63.44**	**62.59**	**62.39**	**62.61**	**61.33**	61.33	61.33	**61.78**	**62.11**	**62.44**	**54.17**	**61.41**	**61.21**	**2.49**	**9.27**
	M-CLIP + MT	**63.44**	61.11	61.33	62.33	61.17	**61.44**	**61.50**	54.83	61.67	58.00	52.72	59.96	59.61	3.35	10.72
Few-shot	M-CLIP	**64.67**	61.83	60.67	61.33	**62.00**	52.22	62.61	**56.33**	60.39	**59.72**	48.61	59.13	58.57	4.83	16.06
	M-CLIP + MT	**64.67**	**61.89**	**62.00**	**62.17**	61.44	**61.00**	**63.00**	56.11	**62.00**	59.50	**53.83**	**60.69**	**60.29**	**3.14**	**10.84**

Table 2: Recall@1 on image-text retrieval dataset XTD for comparison of CLIP and Multilingual-CLIP with machine translation as a tool in zero-shot and few-shot scenarios. We average the score to get the overall performance across the languages, and evaluate the multilingual disparity with standard deviation and range. M-CLIP is short for Multilingual-CLIP and MT is short for machine translation. We bold the best scores for zero-shot and few-shot respectively. "Avg.$_{-en}$" represents average score without English and "en→tgt" means data in other languages is translated from English set.

Setting	XTD			Multi30k		
	Avg.$_{-en}$ ↑	Std↓	Range↓	Avg.$_{-en}$ ↑	Std↓	Range↓
	Few-shot (en: 64.67)			Full (en: 76.10)		
-	58.57	4.83	16.06	74.93	0.70	1.55
MT	60.29	3.14	10.84	75.30	0.68	1.45
Routine1+MSE	58.85	4.72	15.73	75.57	0.72	1.55
Routine2+MSE	58.81	4.75	15.67	75.65	0.95	2.10
Routine3+MSE	**60.52**	**3.11**	**10.50**	**75.97**	**0.60**	**1.45**
Routine3+CL(pivot-tgt)	60.30	3.12	10.73	75.47	0.98	2.15
Routine3+CL(pivot-image)	60.46	3.12	10.56	75.85	0.61	**1.45**

Table 3: We compare different alignment routines for parallel corpus on XTD and Multi30K datasets. We report the score in English once as these combinations cannot applies on English set. CL is short for contrastive loss. We bold the best results on each dataset.

without mastering other languages. For (relatively) high-resource language, the parallel text is also a source to obtain texts in different languages with the same meaning. We call these texts as pivot-target language text pair. Since the model has higher performance on pivot language and those pivot-target text pairs can provide more information, there must be a better approach to exploit pivot language for better cross-lingual transfer of Multilingual-CLIP.

Comparison of different alignment routines and loss functions. In section 3.1, we mention three alignment routines and two alignment loss functions. We compare their different combinations with two baselines without alignment. In the first baseline, we directly apply contrastive loss between images and texts in target language in a mini-batch. For the second baseline, we translate all texts to English previously on the basis of the first baseline. We conduct experiments on XTD and Multi30K in zero-shot, few-shot, and full-dataset learning scenarios. As shown in Table 3, we first compare different routines combined with MSE loss and find that routine 3, translating the target language into English and doing alignment between natural and translation English embedding distribution, performs best. Then we compare different alignments methods with routine 3 and find MSE loss performs best. Ultimately, routine3 combined with MSE loss performs best on all 3 metrics. This can be explained as that Multilingual-CLIP uses MSE loss for text-text pairs in the pre-training stage, and natural English embedding distribution is a better distribution. Meanwhile, machine translation map the target language embedding distribution to a distribution close to optimal distribution where multilingual CLIP performs best, making it easier to optimize.

5.4 Analysis on Parameter-Efficient Cross-Lingual Transfer Learning

In this section, we first evaluate the performance of hard prompt. Then we compare Adapter, Compacter, and LoRA, and discuss the feasibility of using these methods for cross-lingual transfer.

5.4.1 Hard Prompt

We focus on the prompting method, especially the hard prompt method, as it excels at zero-shot learning scenario when domain-specific data is inaccessible for fine-tuning model parameters. In a multilingual scenario, we need to consider two types of prompts. On one hand, prompt might be constructed in a variety of languages, which may result in performance disparity across different languages. On the other hand, the text input can be automatically converted to any other language by machine translation. Therefore, We explore the following combinations: (1) Simply attach prompt in English before text. (2) Translate the best English prompt to target languages and attach them before corresponding texts. (3) Then we translate all the text inputs to English and attach English prompt before them.

Setting	en	de	fr	es	it	ko	pl	ru	tr	zh	jp	Avg.↑	Avg.$_{-en}$ ↑	Std↓	Range↓
Zero-shot															
(1) English Hard Prompt	**64.06**	60.72	59.89	61.78	61.28	51.67	62.06	**56.00**	60.11	**59.78**	47.83	58.65	58.11	4.90	16.23
(2) Target Lang Hard Prompt	**64.06**	60.83	60.33	62.06	61.22	49.00	61.00	**56.00**	59.39	59.33	48.06	58.30	57.72	5.22	16.00
(3) English Hard Prompt + MT	64.06	**61.5**	61.89	**62.28**	**61.56**	60.39	62.39	55.5	**61.89**	58.78	53.22	**60.31**	**59.94**	3.26	10.84
Few-shot															
Soft Prompt (3 tokens)	63.94	61.17	**62.00**	62.17	**61.56**	**60.50**	**62.44**	55.44	61.83	58.72	**53.33**	60.28	59.92	**3.22**	**10.61**
Soft Prompt (20 tokens)	62.72	59.89	60.50	60.28	59.83	59.44	60.22	54.33	60.00	58.11	53.11	58.95	58.57	2.81	9.61

Table 4: Results on XTD dataset for comparison of different combinations of prompts and texts.

Setting	Updated Params per Lang (%)	XTD (few-shot)												Multi30K (Full-dataset)				
		en	de	fr	es	it	ko	pl	ru	tr	zh	jp	Avg.	en	cs	de	fr	Avg.
FT	100	64.67	61.83	60.67	61.33	62.00	52.22	62.61	56.33	60.39	59.72	48.61	59.13	76.10	74.55	74.80	75.45	75.23
Adapter	0.45	64.44	**61.17**	**60.61**	60.72	61.39	52.17	**62.83**	56.44	59.50	59.67	49.17	58.92	75.20	**74.50**	**73.85**	**76.40**	**74.99**
Compacter	0.05	63.67	60.61	60.44	60.50	61.33	52.06	62.67	55.94	59.72	59.39	49.00	58.67	74.25	72.40	73.55	73.85	73.51
LoRA	0.16	**64.83**	61.10	60.39	**60.89**	**61.67**	**53.06**	61.94	**56.67**	**60.50**	**59.82**	**49.56**	**59.13**	75.35	72.30	73.65	74.85	74.04
Our framework	0.16 / 0.45	64.50	**61.94**	**62.00**	**62.00**	61.83	**61.65**	**62.72**	**55.94**	**62.33**	**59.84**	**54.22**	**60.82**	75.35	75.40	**75.15**	**76.40**	**75.58**

Table 5: Comparison of different PET methods. We care about the degree of performance degradation caused by different PET methods. We bold the best score among the three PET methods and bold the score of our framework if it is better than full-model fine-tuning. Our framework updates 0.16% and 0.45% parameters for few-shot and full-dataset scenario, respectively.

Result analysis. From results in Table 4, it can be found that the zero-shot performance increases by simply adding prompt in English, with both text input in target language and translated into English. We compare different hard prompts and find "a photo of" performs best. However, Translating English prompt into target language makes the performance decrease slightly. Finally, We get the best performance by translating all the text inputs into English and adding the best English prompt to them.

Comparison with soft prompt. We also conduct experiments on soft prompt based on the third combination: initiate prompt from the best prompt, which result in 3 trainable token embeddings, and randomly initiate 20 token embeddings. With 50 training instances in the few-shot scenario, the model obtains marginal improvement or even performance decreasing by utilizing these templates. As a result, we do not incorporate soft prompt in our framework.

5.4.2 Other PET Methods

In this section, We compare three popular PET methods, Adapter, Compacter, and LoRA, with full-model fine-tuning. Following He et al. (2022a), we unfreeze the linear head and assign the same learning rate as fine-tuning. The number of parameters of Adapter, Compacter and LoRA is only 0.44%, 0.05% and 0.16% of Multilingual-CLIP's text encode, respectively. Thus, even if we assign different parameters to each of the 100 languages, the total number of parameters is smaller than that of a single model. From the results shown in Table 5, we can find that Adapter performs the best in full-dataset scenario, which preserves 99.7% performance of fine-tuning and LoRA performs the best in few-shot scenario, which even achieve almost the same performance. We further adopt Adapter and LoRA with the best combination discussed in Section 5.3, which forms a part of our final framework, and find this combination obtains better performance than full-model fine-tuning. It indicates that with the help of PET methods, our framework can achieve the parameter-efficiency without losing too much performance.

6 Conclusion

In this paper, we propose a framework that significantly mitigates the multilingual disparity of Multilingual-CLIP in a parameter-efficient manner in zero-shot, few-shot and full-dataset scenarios. Our framework uses a translation-based alignment method and adopts parameter-efficient tuning methods. Analytical experiments indicate that machine translation is effective for cross-lingual transfer; exploiting pivot language can help reduce the disparity; parameter-efficient tuning methods are beneficial for reducing resource consumption without too much performance degradation.

7 Limitations

Our work primarily focuses on addressing multilingual disparity by improving the multilingual text encoder in the CLIP-liked framework. However, it is very possible that the visual encoder can also be

enhanced with image and text data from diverse culture and languages. During the pre-training process, the original Multilingual-CLIP's visual encoder is directly aligned with English corpora only, and connected with other languages by using English as a pivot language. We expect future work can address the multilingual disparity problem from the perspective of a more powerful visual encoder.

Broader Impact

This work provides a framework for cross-lingual transfer in few-shot learning setting. The deployment of our method is potential to mitigate the performance disparity for state-of-the-art multimodal models for scarce-resource languages. However, we note that our method relies on the collection of parallel corpus, either collected from online machine translation systems or native human speakers. Our work does not thoroughly scrutinize whether these parallel corpus contains implicit social biases in different dimensions, such as race, gender and religion. When these parallel corpus contains unexpected biases or stereotypes, it is likely that the model learned from such data may perpetuate these biases that we did not foresee.

References

Pranav Aggarwal and Ajinkya Kale. 2020. Towards zero-shot cross-lingual image retrieval.

Wasi Uddin Ahmad, Zhisong Zhang, Xuezhe Ma, Eduard Hovy, Kai-Wei Chang, and Nanyun Peng. 2018. On difficulties of cross-lingual transfer with order differences: A case study on dependency parsing. *arXiv preprint arXiv:1811.00570*.

Loïc Barrault, Fethi Bougares, Lucia Specia, Chiraag Lala, Desmond Elliott, and Stella Frank. 2018. Findings of the third shared task on multimodal machine translation. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 304–323.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Min Cao, Shiping Li, Juntao Li, Liqiang Nie, and Min Zhang. 2022. Image-text retrieval: A survey on recent research and development. *arXiv preprint arXiv:2203.14713*.

Fredrik Carlsson, Philipp Eisen, Faton Rekathati, and Magnus Sahlgren. 2022a. Cross-lingual and multilingual clip. In *Proceedings of the Language Resources and Evaluation Conference*, pages 6848–6854, Marseille, France. European Language Resources Association.

Fredrik Carlsson, Philipp Eisen, Faton Rekathati, and Magnus Sahlgren. 2022b. Cross-lingual and multilingual clip. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6848–6854.

Xilun Chen, Yu Sun, Ben Athiwaratkun, Claire Cardie, and Kilian Weinberger. 2018. Adversarial deep averaging networks for cross-lingual sentiment classification. *Transactions of the Association for Computational Linguistics*, 6:557–570.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Ning Ding, Yujia Qin, Guang Yang, Fuchao Wei, Zonghan Yang, Yusheng Su, Shengding Hu, Yulin Chen, Chi-Min Chan, Weize Chen, et al. 2022. Delta tuning: A comprehensive study of parameter efficient methods for pre-trained language models. *arXiv preprint arXiv:2203.06904*.

Desmond Elliott, Stella Frank, Loïc Barrault, Fethi Bougares, and Lucia Specia. 2017. Findings of the second shared task on multimodal machine translation and multilingual image description. In *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 215–233, Copenhagen, Denmark. Association for Computational Linguistics.

Desmond Elliott, Stella Frank, Khalil Sima'an, and Lucia Specia. 2016. Multi30k: Multilingual english-german image descriptions. In *Proceedings of the 5th Workshop on Vision and Language*, pages 70–74. Association for Computational Linguistics.

Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. 2021. Clip-adapter: Better vision-language models with feature adapters. *arXiv preprint arXiv:2110.04544*.

Demi Guo, Alexander M Rush, and Yoon Kim. 2020. Parameter-efficient transfer learning with diff pruning. *arXiv preprint arXiv:2012.07463*.

Junxian He, Chunting Zhou, Xuezhe Ma, Taylor Berg-Kirkpatrick, and Graham Neubig. 2021. Towards a unified view of parameter-efficient transfer learning. *arXiv preprint arXiv:2110.04366*.

Xuehai He, Chunyuan Li, Pengchuan Zhang, Jianwei Yang, and Xin Eric Wang. 2022a. Parameter-efficient fine-tuning for vision transformers. *arXiv preprint arXiv:2203.16329*.

Xuehai He, Diji Yang, Weixi Feng, Tsu-Jui Fu, Arjun Akula, Varun Jampani, Pradyumna Narayana, Sugato Basu, William Yang Wang, and Xin Eric Wang. 2022b. Cpl: Counterfactual prompt learning for vision and language models. *arXiv preprint arXiv:2210.10362*.

Geoffrey Hinton, Oriol Vinyals, Jeff Dean, et al. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2(7).

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for nlp. In *International Conference on Machine Learning*, pages 2790–2799. PMLR.

Edward Hu, Yelong Shen, Phil Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models.

Shengding Hu, Zhen Zhang, Ning Ding, Yadao Wang, Yasheng Wang, Zhiyuan Liu, and Maosong Sun. 2022. Sparse structure search for delta tuning. In *Advances in Neural Information Processing Systems*.

Aashi Jain, Mandy Guo, Krishna Srinivasan, Ting Chen, Sneha Kudugunta, Chao Jia, Yinfei Yang, and Jason Baldridge. 2021. Mural: multimodal, multitask retrieval across languages. *arXiv preprint arXiv:2109.05125*.

Rabeeh Karimi Mahabadi, Sebastian Ruder, Mostafa Dehghani, and James Henderson. 2021. Parameter-efficient multi-task fine-tuning for transformers via shared hypernetworks. In *Annual Meeting of the Association for Computational Linguistics*.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021a. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021b. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691*.

Patrick Lewis, Barlas Oğuz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. 2019. Mlqa: Evaluating cross-lingual extractive question answering. *arXiv preprint arXiv:1910.07475*.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Ellie Pavlick, Matt Post, Ann Irvine, Dmitry Kachaev, and Chris Callison-Burch. 2014. The language demographics of Amazon Mechanical Turk. *Transactions of the Association for Computational Linguistics*, 2:79–92.

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.

Afshin Rahimi, Yuan Li, and Trevor Cohn. 2019. Massively multilingual transfer for ner. *arXiv preprint arXiv:1902.00193*.

Andreas Rücklé, Gregor Geigle, Max Glockner, Tilman Beck, Jonas Pfeiffer, Nils Reimers, and Iryna Gurevych. 2020. Adapterdrop: On the efficiency of adapters in transformers. *arXiv preprint arXiv:2010.11918*.

Sebastian Schuster, Sonal Gupta, Rushin Shah, and Mike Lewis. 2018. Cross-lingual transfer learning for multilingual task oriented dialog. *arXiv preprint arXiv:1810.13327*.

Jialu Wang, Yang Liu, and Xin Wang. 2022. Assessing multilingual fairness in pre-trained multimodal representations. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2681–2695, Dublin, Ireland. Association for Computational Linguistics.

Derrick Xin, Behrooz Ghorbani, Ankush Garg, Orhan Firat, and Justin Gilmer. 2022. Do current multi-task optimization methods in deep learning even help? *arXiv preprint arXiv:2209.11379*.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2020. mt5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010.11934*.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *TACL*, 2:67–78.

Elad Ben Zaken, Shauli Ravfogel, and Yoav Goldberg. 2021. Bitfit: Simple parameter-efficient fine-tuning for transformer-based masked language-models. *arXiv preprint arXiv:2106.10199*.

Renrui Zhang, Rongyao Fang, Peng Gao, Wei Zhang, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. 2021. Tip-adapter: Training-free clip-adapter for better vision-language modeling. *arXiv preprint arXiv:2111.03930*.

Yuanhan Zhang, Kaiyang Zhou, and Ziwei Liu. 2022. Neural prompt search. *arXiv preprint arXiv:2206.04673*.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. 2022. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16816–16825.

A Experimental Details

We use Adam (Kingma and Ba, 2014) as our optimizer with a cosine decay learning rate scheduler. For few-shot on XTD dataset, we train the model for 40 epochs and use the batch size of 10, which results in a total of 200 steps. When we train our model on the whole Multi30K, we set the number of epoches to 15 with batch size of 48 due to the memory limitation and the total steps is 9k. When the memory is insufficient, we appropriately reduce batchsize to adapt. We evaluate the performance every 5/300 steps respectively and save the best checkpoint for test. For a more obvious comparison, We report Recall@1 score in all experiments. We froze the image encoder for a fair comparison since the difference between languages is in the text input and tuning image encoder will introduce additional randomness. All the experiments are done on 8 Nvidia V100 GPUs. We use the official code of CLIP and Multilingual-CLIP to load and use pre-training parameters

To find an optimal combination of hyperparameters, we conduct a grid search on learning rate and guidance coefficient λ. The learning rates lie within {3e-5, 1e-4, 3e-4} for parameter-efficient methods, and {1e-6, 3e-6, 1e-5} for fine-tuning the whole model. The guidance coefficients fall within a large range from 0.001 to 10. The optimal λs vary from language to language, but most of them are distributed in a smaller interval from 0.1 to 1.