Efficient Multilingual Multi-modal Pre-training through Triple Contrastive Loss

Youhan Lee
Kakao Brain
youhan.lee@kakaobrain.com

Kyungtae Lim
Hanbat National University
ktlim@hanbat.ac.kr

Woonhyuk Baek, Byungseok Roh, Saehoon Kim
Kakao Brain
wbaek,peter.roh,shkim@kakaobrain.com

Abstract

Learning visual and textual representations in the shared space from web-scale image-text pairs improves the performance of diverse vision-and-language tasks, as well as modality-specific tasks. Many attempts in this framework have been made to connect English-only texts and images, and only a few works have been proposed to extend this framework in multilingual settings with the help of many translation pairs. In this multilingual approach, a typical setup is to use pairs of (image and English-text) and translation pairs. The major limitation of this approach is that the learning signal of aligning visual representation with under-resourced language representation is not strong, achieving a sub-optimal performance of vision-and-language tasks. In this work, we propose a simple yet effective enhancement scheme for previous multilingual multi-modal representation methods by using a limited number of pairs of images and non-English texts. In specific, our scheme fine-tunes a pre-trained multilingual model by minimizing a triple contrastive loss on triplets of image and two different language texts with the same meaning, improving the connection between images and non-English texts. Experiments confirm that our enhancement strategy achieves performance gains in image-text retrieval, zero-shot image classification, and sentence embedding tasks.

1 Introduction

Transferring visual representations learned from a large annotated set into downstream tasks of interest (Deng et al., 2009; Zhai et al., 2019) is the standard approach to achieve the state-of-the-art performance (Kuznetsova et al., 2020; Kolesnikov et al., 2020). However, due to the labeling cost, the scalability of this approach is rather limited. In contrast, self-supervised learning with contrastive losses (He et al., 2020; Chen et al., 2020) has proven to learn semantic representations without using the explicit labels, which becomes a promising solution to obtaining general visual representations. In addition, this paradigm combined with billion-scale unlabeled samples is competitive with the annotation-based transfer learning approaches in multiple tasks (Goyal et al., 2021).

In natural language processing (NLP), pre-training with billions of corpus and transferring to downstream tasks has likewise achieved tremendous success. And it has been a de facto standard for a recent decade (Devlin et al., 2019; Radford et al., 2019; Brown et al., 2020). Witnessing successes in two domains, researchers have been actively attempting to find visual-and-language (VL) representations by combining supervisions that come from billions of both images and languages. (Li et al., 2019; Lu et al., 2019, 2020; Kim et al., 2021) have attempted to utilize highly curated VL datasets such as MS-COCO (Lin et al., 2014) or Visual Genome (Krishna et al., 2016). However, the non-trivial collection process is the major limitation in scaling up these datasets. CLIP (Radford et al., 2021) overcame the limitation by learning VL representations from web-scale 400M image-to-text pairs on both visual and VL downstream tasks, even with a simple contrastive loss. ALIGN (Jia et al., 2021) extended this approach by scaling up image-text pairs to 1.8B with simple filtering compared to CLIP and showed state-of-the-art scores on both visual and visual-language tasks.

Such web-scale approaches as well as the earliest attempts have focused on connecting images to English texts. Using a separate translator might be a practical solution to match images and multilingual texts, but this resorts to sub-optimal results because the effect of translation errors is not explicitly considered in the VL representation learning process, and translating every text into dozens of languages is inefficient. In language models, many attempts have been successfully made to develop language-agnostic representations, improving the performance of downstream tasks on under-

Proceedings of the 29th International Conference on Computational Linguistics, pages 5730–5744
October 12–17, 2022.

resourced languages. In order to achieve this goal, Pires et al. (2019); Liu et al. (2020); Xue et al. (2021) pre-train language models over a multilingual corpus for transferring knowledge across languages. Chi et al. (2021); Feng et al. (2020) utilize bilingual translation pairs to transfer the information from common languages to under-resourced ones in a more efficient way.

Following this research direction in language models, MURAL (Jain et al., 2021) suggests the efficient multilingual VL modeling that leverages both the alignments of (a) monolingual image-to-text and (b) multilingual text-to-text on the training dataset simultaneously. However, since English is the only language that has a direct connection with images, the performance of non-English languages on several multi-modal tasks is relatively weaker than that of English. Linking all languages directly to images would be the easiest and ideal solution, but obtaining billions of image-text datasets for all languages is nearly impossible, especially for low-resourced languages.

To tackle the data-scarcity problem, we propose a simple but efficient enhancement via triplet contrastive learning (ETCL) that utilizes multi-modal zero-shot transfer through relatively small amounts of image-text datasets in non-English languages through a triplet contrastive loss. In ETCL, we further train the pre-trained multilingual VL model with a new limited number of pairs of image and non-English texts to strengthen the weak alignment of the pre-trained model. In order to fully leverage multi-modal cross-lingual zero-shot transfer, we introduce a triplet contrastive learning which considers (a) image-textA, (b) textA-textB and (c) textB-image, where textA and textB are multilingual translation pairs.

Through various experiments, we show that our framework provides significant performance improvement for various languages in multi-modal retrieval tasks and zero-shot image classification and sentence embedding tasks. Interestingly, the large performances gains also occur in some languages which are not included in the training dataset for ETCL. Through ablation study, we prove that the phenomenon comes from the multi-modal cross-lingual zero-shot transfer with regard to grammatical and geographic relationships between languages. And we show that ETCL leverages the relationships efficiently. In summary, our contributions to the multilingual VL representation are:

- We propose the ETCL framework that leverages multi-modal cross-lingual zero-shot transfer to train VL model efficiently.
- We show that ETCL gives the large performance gains on various downstream tasks such as image-text retrieval, zero-shot image classification and language tasks.
- We provide the empirical analysis that ETCL utilizes the geographical and grammatical relationships between languages efficiently.

2 Related works

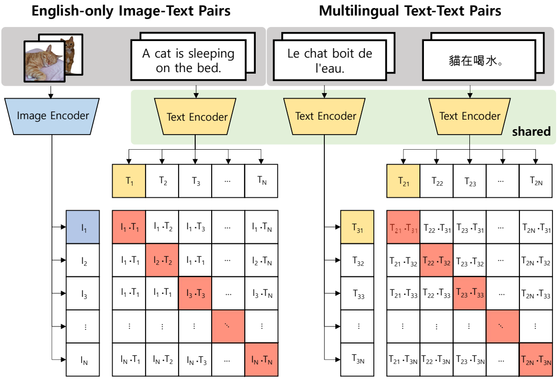
We briefly review the research areas and key references most relevant to our approach.

2.1 Multilingual representation learning

Multilingual language models have shown that a single large model improves diverse multilingual NLP tasks, removing the need for maintaining language-specific models (Pires et al., 2019; Liu et al., 2020; Xue et al., 2021; Feng et al., 2020; Chi et al., 2021). Multilingual BERT (Pires et al., 2019), dubbed as mBERT, is a multilingual variant of BERT (Devlin et al., 2019), which is pre-trained with the masked language modeling objective over about 100 languages. LaBSE (Feng et al., 2020) adapts mBERT to learn language-agnostic sentence embedding over bilingual translation pairs, improving cross-lingual retrieval tasks significantly. mBART (Liu et al., 2020) trains an encoder-decoder architecture on a large corpus composed of multiple languages using BART objective (Lewis et al., 2019). Considering multilingual variants to newly proposed models has still been actively studied. For instance, T5 (Raffel et al., 2020) is also extended to a multilingual model, which is named mT5 (Xue et al., 2021), by training a sequence-to-sequence model over 101 languages. Recently, mT6 (Chi et al., 2021) further improves mT5 by proposing a novel objective for text-to-text pre-training.

2.2 Vision-and-Language representation learning

Learning VL representations in a self-supervised fashion is a promising approach to solving many visually grounded language understanding tasks (Li et al., 2019; Lu et al., 2019, 2020; Kim et al., 2021). ViLBERT (Lu et al., 2019) and VisualBERT (Li et al., 2019) try to align patches in an


English-only Image-Text Pairs
Multilingual Text-Text Pairs
A cat is sleeping on the bed.
Le chat boit de l'eau.
貓在喝水。
Image Encoder
Text Encoder
Text Encoder
Text Encoder
shared
T1 T2 T3 ... TN
I1 I2 I3 ⋮ IN
I1·T1 I1·T2 I1·T3 ... I1·TN
I1·T1 I2·T2 I1·T1 ... I2·TN
I1·T1 I1·T1 I3·T3 ... I3·TN
IN·T1 IN·T2 IN·T3 ... IN·TN
T21 T22 T23 ... T2N
T31 T32 T33 ⋮ T3N
T21·T31 T22·T31 T23·T31 ... T2N·T31
T21·T32 T22·T32 T23·T32 ... T2N·T32
T21·T33 T22·T33 T23·T33 ... T2N·T33
T21·T3N T22·T3N T23·T3N ... T2N·T3N

(a) Image-to-Text (I2T) Contrastive Loss

(b) Text-to-Text (T2T) Contrastive Loss

Figure 1: Illustration on English-only image-to-text and multilingual text-to-text matching, where the text encoder is shared across languages.

image into (sub-)words through a sequence of self-attention layers, generating transferable representations on many VL downstream tasks. ViLT (Kim et al., 2021) advances the model architecture and training procedure to achieve comparable performance without using the region proposal network. Since these works rely on highly curated multi-modal datasets, the scalability of these approaches is rather limited. CLIP (Radford et al., 2021) and ALIGN (Jia et al., 2021) have shown that it is possible to learn transferrable visual and VL representations from noisy web-scale datasets. Very recently, MURAL (Jain et al., 2021) extends ALIGN in a multilingual setting by leveraging billion-scale translation pairs. However, the performance of under-resourced languages on various VL tasks is still weaker than that of English. Otherwise, KELIP(Ko and Gu, 2022), a CLIP-style bilingual VL model trained using 708M Korean and 476M English image-text pairs, has been proposed, showing strong representations on Korean VL tasks. However, the approach is not an optimal solution for multilingual VL modeling because gathering image-text pairs for more than a hundred languages is practically impossible. Our work suggests a simple but effective solution to boost the performance of under-resourced languages.

3 Approach

This section describes the details of conventional multilingual VL modeling, the proposed ETCL scheme, and training details.

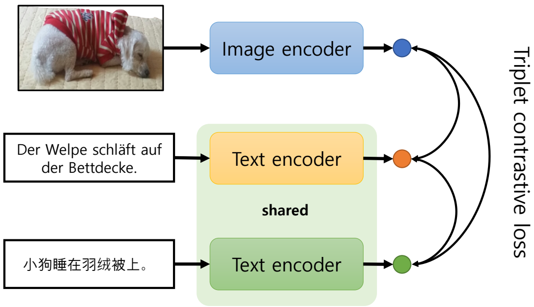


Figure 2: Illustration on a triple contrastive loss using (1) image-textA, (2) textA-textB and (3) textB-image, where the text encoder is shared across languages.

3.1 Background: Multilingual VL pretraining

MURAL (Jain et al., 2021) suggests large-scale multilingual VL modeling using pairs of (image and English-text) and translation pairs. (see Figure.1). For training VL representations, the system requires capturing the meaning of images and texts at the same time. In this context, contrastive loss is a suitable objective since it places semantically similar images and sentences in the same latent vector space as follows:

$$\begin{aligned}\mathcal{L}_{\text{i2t}} &= -\frac{1}{N}\sum_{i=1}^{N}\frac{e^{\phi(x_i,y_i)}}{\sum_{n=1}^{N}e^{\phi(x_i,y_i)}},\\ \mathcal{L}_{\text{t2i}} &= -\frac{1}{N}\sum_{i=1}^{N}\frac{e^{\phi(y_i,x_i)}}{\sum_{n=1}^{N}e^{\phi(y_i,x_i)}},\end{aligned} \tag{1}$$

$$\mathcal{L}_{\text{image-text}} = \mathcal{L}_{\text{i2t}} + \mathcal{L}_{\text{t2i}}, \tag{2}$$

where $\mathcal{L}_{\text{i2t}}$ and $\mathcal{L}_{\text{t2i}}$ represent the image-to-text and text-to-image losses respectively and ϕ denotes scoring function $x^{\top}y$. Also, x_i and y_i denote the image and text features in the i-th pair, N is the batch size, σ is a learnable softmax temperature. The loss function for text-to-text matching from translation pairs is similarly defined and denoted by $\mathcal{L}_{\text{textA-textB}}$ and $\mathcal{L}_{\text{textB-textA}}$, and $\mathcal{L}_{\text{text-text}}$ is defined as the sum of two losses. The final objective of MURAL is followed by:

$$\mathcal{L}_{\text{MURAL}} = \frac{1}{2}\left(\mathcal{L}_{\text{image-text}} + \mathcal{L}_{\text{text-text}}\right), \tag{3}$$

In addition to this, MURAL adds an addictive margin (Yang et al., 2019) to $\mathcal{L}_{\text{MURAL}}$.

3.2 Enhancement via triple contrastive loss

Although MURAL trains a multilingual VL representation efficiently, the alignment of visual representation with under-resourced language representation is still weaker than that of English. To tackle this limitation, we suggest an enhancement scheme that conducts training by adding a small amount of non-English image-text dataset to improve weak alignment between image and non-English languages. As with (Ruan et al., 2022), we introduce a triple contrastive loss that takes into account all pair possibilities rather than simply training with $\mathcal{L}_{\text{image}-\text{text}}$ between images and non-English captions to take full advantage of zero-shot transfers across multi-modal cross-lingual in the training dataset (see Figure 2). Before defining the proposed triple contrastive loss, we prepare image, and two pairs of text in different languages with the same meaning. Since we have prepared three different representations of training data with the same semantic, we can simply apply contrastive loss to the three inputs(image, textA, textB) as follows:

$$\mathcal{L}_{\text{ETCL}} = \frac{1}{3}\left(\mathcal{L}_{\text{image-textA}} + \mathcal{L}_{\text{textA-textB}} + \mathcal{L}_{\text{textB-image}}\right). \quad (4)$$

Here, we refer to our enhancement scheme as ETCL, abbreviated version of **E**nhancement via **T**riplet **C**ontrastive **L**earning. Combined with MURAL, we name it as MURAL+ETCL.

3.3 Model details

Since there is no publicly available MURAL model, we reproduce the MURAL using the architecture and training method as proposed in MURAL (Jain et al., 2021). The image-text pairs required are created by ourselves, and the publicly available CCMatrix is used for text-text pairs.

Model architecture Our model is composed of an image encoder and three text encoders to align image-text and text-text pairs (see Figure 1), where the parameters of all text encoders are shared. For the image encoder, we follow the same protocol of ALIGN (Jia et al., 2021) by using EfficientNet (Tan and Le, 2019) with global averaging pooling to obtain the embedding of an image. For the text encoder, we choose BERT (Devlin et al., 2019) and use the hidden representation of [CLS] token as the embedding of a text. To make sure that the embeddings have the same dimension, we add an additional fully-connected layer on the top of the text encoder.

Pre-processing For a fair comparison, we try to follow the data pre-processing and augmentation used in ALIGN and MURAL as much as possible. For training, we resize images of arbitrary resolutions into 346 × 346 resolution regardless of the aspect ratio. Then, we randomly crop the image to 289 × 289, and apply the horizontal flip of probability 0.5. For evaluation, we resize the image to 346 × 346, and apply the center crop of 289 × 289. For both training and evaluation, we use the same pre-processing for the texts. We use the tokenizer having 550K vocabulary provided by the official repository of LaBSE [1], and truncate the sequence to have the maximum length of 64.

Training details For MURAL optimization, we use LAMB optimizer (You et al., 2020) with a weight decay ratio of 1e-5. The learning rate is linearly increased from zero to 1e-3 in 10k steps, and then linearly decays to zero in 800k steps. We use the label smoothing of 0.1. The temperature is initialized as 1.0, and the margin m is 0.3 same as to LaBSE (Feng et al., 2020). We train the model with a batch size of 16,384 on 128 Cloud TPU V3 cores with 128 positive pairs on each core. Since a large number of negative samples is critical in contrastive learning, we adopt the cross-accelerator negative sampling as used in LaBSE. This enables the large batch training by collecting samples in all synchronized cores and treating them as negative samples. For ETCL, we use a 32,768 batch size with a learning rate of 1e-4, which decreases from 1e-4 to zero linearly in 3k steps.

3.4 Data collection

Image-text pairs We follow the data collection process of ALIGN to create our in-house 1B image-text pairs dataset from Common crawl [2]. The raw descriptions are gathered from the Alt-text HTML attribute associated with web images. We only apply minimal rule-based filtering as detailed below. For image-based filtering, we only keep images whose shorter dimension is larger than 200 pixels and set the aspect ratio as 3. We discard images with more than 100 associated alt-text. For text-filtering, we exclude alt-texts that are shared by more than 10 images. We also discard either too short (<3 unigrams) or too long (>100 unigrams or >1000 characters). These filters include discarding instances that are classified as non-English by

[1] https://tfhub.dev/google/LaBSE/2

[2] https://commoncrawl.org/the-data/

cld3 (compact language detector v3) [3]. Additionally, we also include CC3M (Sharma et al., 2018) and CC12M (Changpinyo et al., 2021) in our 1B image-text pairs dataset.

Text-text pairs We use all 10.8B text-text pairs in CCMatrix dataset (Schwenk et al., 2021) for the text-test alignment, which covers 90 languages and have 1,197 bitexts across multiple languages. All pairs are publicly available on the OPUS website [4].

Multilingual image-text pairs We use multilingual CC3M datasets which consist of translated versions of original CC3M in 5 languages (de, fr, cs, zh and ja), which are provided (Zhou et al., 2021). After generating 15 language pairs in 6 languages without duplicates, image-textA-textB 45M triple dataset was created and used for `ETCL`. In this dataset, the number of unique images is 3M, and the captions are 18M.

4 Experiments

In this section, we present experiment settings to evaluate our proposed model over VL, visual, and language tasks.

4.1 Task description

Image-text retrieval The image-text retrieval is the most suitable task to evaluate the VL model because this task uses the representation of language and image simultaneously. Following (Jain et al., 2021; Jia et al., 2021; Li et al., 2019), we validate the ability of our VL representation on multilingual Flickr30K (Young et al., 2014; Elliott et al., 2016, 2017; Barrault et al., 2018) and MS-COCO (Lin et al., 2014; Yoshikawa et al., 2017) in image-to-text and text-to-image retrieval tasks with zero-shot and fine-tuning scenarios. In this experiment, we use multilingual Flickr30K having English, German, French, and Czech captions, where and German captions are provided by Multi30K (Elliott et al., 2016) and French and Czech captions are obtained by the translation (Elliott et al., 2017; Barrault et al., 2018). Multi30K contains five captions per image in English and German, and one description per image in French and Czech. We use 29K, 1K, and 1K images for the train, validation, and test sets, respectively as used in the original dataset (Young et al., 2014). In addition to English MS-COCO, Japanese (Yoshikawa et al., 2017) and Korean MS-COCO [5] are also used for the VL evaluation. We follow the split protocol used in (Karpathy and Fei-Fei, 2015), resulting in 82K training and 5K test sample. For evaluation, we measure Recall@K with respect to $K = 1, 5, 10$ on two retrieval tasks. We report the performance of each model by Average Recall (AR), taking the mean over these six scores.

During fine-tuning, we follow the same protocol used in ALIGN for a fair comparison. The pre-trained model is fine-tuned by the image-text contrastive loss (without text-text matching loss). We use the batch size of 2,048 and set the learning rate to 1e-5 with a linear decay scheduler, and fine-tune the model over 3K and 6K steps on Flickr30K and MS-COCO, respectively. All the other hyper-parameters are consistent with the ones in pre-training.

Zero-shot image classification Performance on the zero-shot image classification has been considered one of the important evaluation tasks for the large-scale multi-modal pre-training model since it represents the generalization ability of a model. Following previously proposed studies(Jain et al., 2021; Jia et al., 2021), we validate the visual representation power of our method on multilingual zero-shot classification tasks based on text prompts. Furthermore, an additional result on ImageNet K-Nearest-neighbor (KNN) tasks without text prompts is provided in Appendix A.3. Our models are evaluated on diverse classification datasets, including ImageNet (Deng et al., 2009), CIFAR100 (Krizhevsky, 2009), SUN397 (Xiao et al., 2010), and Fool101 (Bossard et al., 2014). We conduct the zero-shot image classification based on text prompts (Radford et al., 2021) over 83 languages, where prompts are translated by Google Translator. We remark that the translator supports 83 languages among 90 languages covered in CCMatrix. For a fair comparison, we adopt the text prompt engineering used in CLIP. For instance, in the case of Food101, a context prompt is inserted, so the final prompt is *“A photo of a {label}, a type of food"*, and the context prompt is also translated over 83 languages. Similar to MURAL, we compare models on three groups, `All-languages`, `well-resourced` and `under-resourced` to deeply investigate our model in different resource conditions. The

[3] `https://github.com/google/cld3`
[4] `https://opus.nlpl.eu/CCMatrix.php`
[5] Korean COCO is from `https://aihub.or.kr/keti_data_board/visual_intelligence`.

Type	Model	Backbone	Image-text	Text-text	Flickr30K					MSCOCO 5K			
			pairs	pairs	en	de	fr	cs	avg.	en	ja	ko	avg.
Zero-shot	ALIGN-BASE	B5+BERT-Base	1.8B	-	83.3	75.0	74.2	47.9	70.1	59.6	53.9	-	-
	ALIGN-L2	L2+BERT-Large			**92.2**	-	-	-	-	**70.9**	-	-	-
	MURAL-BASE	B5+BERT-Base	1.8B	6B	82.4	76.2	75.0	64.6	74.6	59.5	54.4	-	-
	MURAL-LARGE	B7+BERT-Large			89.2	83.5	83.1	77.0	83.2	67.7	**64.6**	-	-
	MURAL(reprod.)	B7+BERT-Base	1B	10.8B	90.1	70.5	70.1	63.7	73.6	67.8	40.6	31.5	46.6
	MURAL(reprod.) + ETCL	B7+BERT-Base	1B + 45M	10.8B + 45M	90.6	**85.6**	**85.8**	**82.2**	**86.0**	69.1	62.1	**49.3**	**60.2**
Fine-tune	ALIGN-BASE	B5+BERT-Base	1.8B	-	92.2	88.5	88.1	84.5	88.3	74.8	72.5	-	-
	ALIGN-L2	L2+BERT-Large			**96.0**	-	-	-	-	**83.4**	-	-	-
	MURAL-BASE	B5+BERT-Base	1.8B	6B	92.2	88.6	87.6	84.2	88.2	75.4	74.9	-	-
	MURAL-LARGE	B7+BERT-Large			93.8	90.4	89.9	87.1	90.3	81.2	**81.3**	-	-
	MURAL(reprod.)	B7+BERT-Base	1B	10.8B	94.5	89.8	90.2	87.8	90.6	79.3	77.9	73.3	76.9
	MURAL(reprod.) + ETCL	B7+BERT-Base	1B + 45M	10.8B +45M	94.8	**91.0**	**91.3**	**89.7**	**91.7**	79.8	79.2	**77.0**	**77.8**

Table 1: Average Recall of image-to-text and text-to-image retrieval tasks on multilingual Flickr30K and MS-COCO in zero-shot and fine-tuning scenarios for English (en); German (de); French (fr); Czech (cs); Japanese (ja). The numbers of ALIGN and MURAL are taken from (Jain et al., 2021). The last column for each dataset means the average score over all languages.

groups are below:

- well-resourced: English (en), German(de), French (fr), Czech (cs), Japanese (ja), Chinese (zh), Russian (ru), Polish (pl), Turkish (tr)
- under-resourced: Uzbek (uz), Irish (ga), Belarusian (be), Malagasy (mg), Cebuano (ceb)

Sentence similarity & retrieval For the multilingual VL model, one important question is still remaining: *what does the model learn from languages and how can we assess its ability to NLP tasks?* As previous VL studies proposed (Jain et al., 2021), we validate the performance of sentence embeddings from our approach in (multilingual and monolingual) sentence similarity comparison (Cer et al., 2017) and cross-language sentence retrieval tasks (Artetxe and Schwenk, 2019). For the sentence similarity comparison, we choose STS 2017 (Cer et al., 2017) and its extended version (Reimers and Gurevych, 2020). Both datasets contain human-annotated labels of the similarity from 0 (no meaning overlapping) to 5 (equivalent meaning) for every pair of sentences. We compute the Spearman rank correlation between the cosine similarity of two sentence embeddings and the ground-truth label in both 3 monolingual and 7 multilingual settings. We also evaluate ETCL on a multilingual sentence retrieval task with Tatoeba dataset (Artetxe and Schwenk, 2019). In this experiment, we observe that ETCL have decent performance compared to other language models, and more details could be found in Appendix A.5.

5 Results

As presented in our contributions in Section 1, our study has different goals: (1) investigate the performance gains based on the proposed method on various downstream tasks including image-text retrieval, zero-shot image classification and sentence retrieval (2) study the effect of ETCL in related languages (geographically and grammatically), (3) check whether ETCL activates information sharing in multilingual training. By taking into account the proposed goals, in this section, we report our experimental results in various tasks.

5.1 Downstream Task Results on Image-Text Retrieval

To investigate the effectiveness of the proposed model, we conducted experiments on the Image-Text Retrieval task. Table 1 shows the average recall over three different Ks of two retrieval tasks in zero-shot and fine-tuning scenarios. We note that the reproduction of MURAL is successful because MURAL(reprod.) performs comparably to ALIGN-BASE and MURAL-LARGE in all languages of the two datasets.

First, for Flickr30K, ALIGN-L2 performs better than MURAL(reprod.)+ETCL in both zero-shot and fine-tuning cases (92.2 vs. 90.1 and 96.0 vs. 94.5) in the case of English, because the capacity of the image encoder of ALIGN-L2 is larger than MURAL(reprod.)+ETCL. In the case of non-English, we observe that MURAL(reprod.)+ETCL model shows better performance compared to MURAL-LARGE and ALIGN-L2 on German, French in zero-shot and German, French, and Czech in fine-tune setting even if the model capacity is smaller than that of two models.

For MS-COCO, MURAL(reprod.)+ETCL also shows competitive performance for all languages in both zero-shot and fine-tune cases with the relatively small size of model compared to MURAL-LARGE and ALIGN-L2. Especially for Korean, which is not included in languages of new image-text pairs for ETCL, MURAL(reprod.)+ETCL outperforms MURAL(reprod.) in both zero-shot and fine-tuning cases by 17.8 and 3.7, respectively. These results confirm that multi-modal cross-lingual zero-shot transfer occurs effectively for languages not in training data after ETCL. In other words, zero-shot transfer strengthens the weak connection between image and non-English text after indirect alignment training. A more in-depth study of the phenomenon will be discussed in Section 5.4.

In Appendix A.1, all Recall@K with respect to $K = 1, 5, 10$ on two retrieval tasks for all languages are provided.

5.2 Downstream Task Results on Zero-Shot Image Classification

Language	Model	ImageNet	SUN397	Food101	CIFAR100
All languages	MURAL(reprod.)	18.4	29.6	29.1	16.6
	MURAL(reprod.) + ETCL	**26.9**	**37.1**	**45.1**	**19.6**
Well-resourced	MURAL(reprod.)	26.4	40.4	29.1	16.6
	MURAL(reprod.) + ETCL	**38.1**	**50.7**	**45.1**	**19.6**
Under-resourced	MURAL(reprod.)	15.2	23.0	29.4	16.7
	MURAL(reprod.) + ETCL	**23.6**	**29.2**	**45.5**	**19.8**

Table 2: Classification accuracy (%) on multilingual zero-shot image classification on 4 datasets. Languages are grouped as followed (1) All languages, (2) well-resourced languages and (3) under-resourced languages group. All values are the average of 100 × Acc@1 across languages.

As presented in Section 4.1, we conduct zero-shot image classification and languages are grouped to see general trends depending on language resources. Table 2 shows the zero-shot classification performance on four datasets. All detailed results are illustrated in Appendix A.4. We note that all results are reliable to compare the trends because MURAL(reprod.)+ETCL shows comparable performance compared to CLIP-ViT/32 (62.6 vs 64.6). For all three groups, MURAL(reprod.)+ETCL shows better performance on four datasets. For the well-resourced group, the performance of ru, pl and tr improves although those languages are not in the newly added languages. Above all, the performance of the under-resourced group which does not include any newly added languages also improves largely over all datasets, which confirms that ETCL gives effective multi-modal zero-shot cross-lingual transfer over low-resourced languages as well.

5.3 Results on language tasks

To answer the raised question about assessing our model in NLP tasks in Section 4.1, we conduct NLP-oriented experiments. The goal of this experiment is to verify how well sentence embeddings obtained from pre-trained models can solve sentence similarities and search tasks.

	en-en	es-es	ar-ar	Avg.
LaBSE	79.4	80.8	69.1	76.4
MURAL(reprod.)	84.8	80.5	69.2	78.1
MURAL(reprod.) + ETCL	**87.3**	**83.3**	**76.2**	**82.3**

Table 3: Performance on extended STS 2017 similarity comparison task in the monolingual setting. Scores are calculated by 100 × Spearman rank correlation between the cosine similarity of sentence embeddings and the gold labels.

	en-ar	en-de	en-tr	en-es	en-fr	en-it	en-nl	Avg.
LaBSE	74.5	73.8	72.0	65.5	77.0	76.9	75.1	73.5
MURAL(reprod.)	70.7	70.4	69.8	69.1	72.9	71.8	72.4	71.0
MURAL(reprod.) + ETCL	**75.7**	**84.1**	**76.3**	**79.5**	**85.3**	**82.1**	**82.3**	**80.7**

Table 4: Performance on extended STS 2017 similarity comparison task in the multilingual setting. Scores are calculated by 100 × Spearman rank correlation between the cosine similarity of sentence embeddings and the gold labels.

Semantic Textual Similarity To evaluate the sentence embedding performance of the text encoder, the STS task was performed with the same evaluation protocol as done in (Reimers and Gurevych, 2019). Since the text encoder is trained with a contrastive loss, we set LaBSE, a text-encoder-only language model trained in the same way, as a baseline. As shown in Table 3 and 4, the performance of the text encoder of MURAL(reprod.) is similar to that of LaBSE. This is because the two text encoders are trained using text-text alignment based on the parallel text pairs. Interestingly, we can see that the performance of MURAL(reprod.)+ETCL increases significantly by 4.2 and 9.7 in STS 2017 monolingual and multilingual settings respectively. We conjecture that this performance improvement is due to the synergy caused by the image domain because the multi-modal model is trained with both images and text simultaneously. In summary, these results reveal that since the text encoder of the VL model is trained along with the image encoder, there is room for improvement in the performance of the text encoder by using the image encoder, unlike the text encoder-only language model. a

Model	Avg	Gain over baseline
MURAL(reprod.) baseline	31.5	-
MURAL(reprod.) + cs	47.1	15.6
MURAL(reprod.) + fr	45.4	13.9
MURAL(reprod.) + ja	47.6	16.0
MURAL(reprod.) + zh	47.1	15.6
MURAL(reprod.) + de	46.1	14.6

Table 5: Performance on Korean MS-COCO image-text and text-image retrieval tasks. Scores are the average of T2I and I2T R@1,5,10.

Model	Avg	Gain over baseline
MURAL(reprod.) baseline	40.6	-
MURAL(reprod.) + cs	54.3	13.7
MURAL(reprod.) + fr	53.2	12.6
MURAL(reprod.) + ja	52.5	11.9
MURAL(reprod.) + zh	52.3	11.7
MURAL(reprod.) + de	52.4	11.8

Table 6: Performance on Ukraine MS-COCO image-text and text-image retrieval tasks. Scores are the average of T2I and I2T R@1,5,10.

Model	3 COCO datasets	4 Flickr30K datasets
MURAL(reprod.)	139.9	294.4
MURAL(reprod.) + 6 lang	179.3	343.9
MURAL(reprod.) + ETCL	180.5	344.2

Table 7: Performance on MS-COCO(en, ja, ko) and Flickr30K(en, de, fr, cs) in zero-shot image-text retrieval tasks. Scores are the summation of averages of 6 scores (T2I and I2T R@1,5,10).

5.4 Analysis and ablation study

Observing the overall performance in Table 1, one could be convinced that adding more multilingual image-text pairs during `ETCL` will bring better performance. Then, one can ask some questions (1) does our model really transfer VL knowledge even for unseen language?, (2) how does zero-shot transfer differ depending on linguistic context?. Further experiments and analyses are performed to gain a better understanding of multi-modal zero-shot transfer. In addition, we investigate the effectiveness of triple contrastive loss. Another ablation study on VL pre-training steps is described in Appendix A.3.

Effect of the multi-modal zeroshot-transfer In order to answer questions (1) and (2), we fine-tune `MURAL(reprod.)` with $\mathcal{L}_{\text{image-text}}$ using one language pairs from multilingual CC3M and compare models on the Korean and Ukraine [6] COCO image-text retrieval. As shown in rows 2-6 in both Table 5 and Table 6, additional training using ($\mathcal{L}_{\text{image-text}}$) brings zero-shot transfer gain over both unseen Korean and Ukraine. Interestingly, additional training using Japanese image-text pairs shows the largest gain over other languages for Korean COCO. We conjecture that Japanese is a language very similar to Korean grammatically, resulting in enhanced zero-shot transfer. Likewise, Czech, similar to Ukraine geographically, also has the largest zero-shot transfer gain in Ukraine COCO. Another ablation study is included in Appendix A.5.

Effectiveness of triple contrastive loss One can expect that the cross-lingual zero-shot transfer effect can sufficiently occur with only the $\mathcal{L}_{\text{image-text}}$. To investigate the effect of the triple contrastive loss, we fine-tune `MURAL(reprod.)` with $\mathcal{L}_{\text{image-text}}$ using all six multilingual CC3M (we name it `MURAL(reprod.)+6lang`). Furthermore, we compare it with `MURAL(reprod.)+ETCL` on zero-shot image-text retrieval for multilingual MS-COCO (en, ja, ko) and Flickr30K (en, de, fr, cs). As shown in Table 7, row 1-2 show that cross-lingual zero-shot transfer also occurs in `MURAL(reprod.)+6lang` as expected. However, `MURAL(reprod.)+ETCL` still show the best performance, proving the effectiveness of triple contrastive loss.

6 Conclusion

VL modeling using large-scale web-crawled datasets has shown great success but cannot be easily utilized for low-resourced languages due to the lack of data. Although a method to efficiently create a multilingual VL model through indirect text-text alignment has been proposed, the VL representations of low-resourced languages are still

[6] We translate MS-COCO into Ukraine using google translator.

weaker than that of English. This work proposes a new approach to solve the limitation based on the observation that multi-modal zero-shot transfer occurs with regard to grammatical and geographical relationships between languages. Our proposed method can be easily adapted to multilingual multi-modal models trained similarly to MURAL.

7 Ethical consideration

It is likely that our model has unwanted social bias from the majority of English in our dataset. If an objective or a dataset that strongly increases the contribution of low-resource language is complemented, the bias can be alleviated.

In our approach, a larger model size, more data, and longer training steps lead to better models. However, these lead to an environmental impact inevitably. Therefore, more research is required to develop a large-scale multilingual multi-modal model with fewer steps and a small backbone to alleviate tremendous computing resources harming our environment.

References

Mikel Artetxe and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Trans. Assoc. Comput. Linguistics*.

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*.

Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. 2014. Food-101 – mining discriminative components with random forests.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, T. J. Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeff Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *ArXiv*, abs/2005.14165.

Daniel Matthew Cer, Mona T. Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. Semeval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *SemEval@ACL*.

Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. 2021. Conceptual 12M: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *CVPR*.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. 2020. A simple framework for contrastive learning of visual representations.

Zewen Chi, Li Dong, Shuming Ma, Shaohan Huang, Xian-Ling Mao, Heyan Huang, and Furu Wei. 2021. mt6: Multilingual pretrained text-to-text transformer with translation pairs.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding.

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*.

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*.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2020. Language-agnostic BERT sentence embedding. *CoRR*, abs/2007.01852.

Priya Goyal, Mathilde Caron, Benjamin Lefaudeux, Min Xu, Pengchao Wang, Vivek Pai, Mannat Singh, Vitaliy Liptchinsky, Ishan Misra, Armand Joulin, and Piotr Bojanowski. 2021. Self-supervised pretraining of visual features in the wild. *arXiv*.

Jiyeon Ham and Eun-Sol Kim. 2021. Semantic alignment with calibrated similarity for multilingual sentence embedding. In *Findings of the Association for Computational Linguistics: EMNLP 2021*.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross B. Girshick. 2020. Momentum contrast for unsupervised visual representation learning.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalization. *CoRR*.

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*.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *ICML*.

Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3128–3137.

Wonjae Kim, Bokyung Son, and Ildoo Kim. 2021. Vilt: Vision-and-language transformer without convolution or region supervision.

Byungsoo Ko and Geonmo Gu. 2022. Large-scale bilingual language-image contrastive learning. *arXiv preprint arXiv:2203.14463*.

Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Joan Puigcerver, Jessica Yung, Sylvain Gelly, and Neil Houlsby. 2020. Large scale learning of general visual representations for transfer.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, Michael Bernstein, and Li Fei-Fei. 2016. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *arXiv*.

Alex Krizhevsky. 2009. Learning multiple layers of features from tiny images.

Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper R. R. Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Tom Duerig, and Vittorio Ferrari. 2020. The open images dataset V4: unified image classification, object detection, and visual relationship detection at scale.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. Visualbert: A simple and performant baseline for vision and language. *arXiv*.

Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. 2014. Microsoft coco: Common objects in context.

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.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks.

Jiasen Lu, Vedanuj Goswami, Marcus Rohrbach, Devi Parikh, and Stefan Lee. 2020. 12-in-1: Multi-task vision and language representation learning.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *ArXiv*, abs/1908.10084.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation.

Yue Ruan, Han-Hung Lee, Ke Zhang, and Angel X Chang. 2022. Tricolo: Trimodal contrastive loss for fine-grained text to shape retrieval. *arXiv preprint arXiv:2201.07366*.

Holger Schwenk, Guillaume Wenzek, Sergey Edunov, Edouard Grave, Armand Joulin, and Angela Fan. 2021. CCMatrix: Mining billions of high-quality parallel sentences on the web.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning.

Mingxing Tan and Quoc V. Le. 2019. Efficientnet: Rethinking model scaling for convolutional neural networks.

Jianxiong Xiao, James Hays, Krista A. Ehinger, Aude Oliva, and Antonio Torralba. 2010. Sun database: Large-scale scene recognition from abbey to zoo.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mt5: A massively multilingual pre-trained text-to-text transformer.

Yinfei Yang, Gustavo Hernandez Abrego, Steve Yuan, Mandy Guo, Qinlan Shen, Daniel Cer, Yun hsuan Sung, Brian Strope, and Ray Kurzweil. 2019. Improving multilingual sentence embedding using bidirectional dual encoder with additive margin softmax.

Yinfei Yang, Daniel Cer, Amin Ahmad, Mandy Guo, Jax Law, Noah Constant, Gustavo Hernandez Abrego, Steve Yuan, Chris Tar, Yun-hsuan Sung, Brian Strope, and Ray Kurzweil. 2020. Multilingual universal sentence encoder for semantic retrieval. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

Yuya Yoshikawa, Yutaro Shigeto, and Akikazu Takeuchi. 2017. Stair captions: Constructing a large-scale japanese image caption dataset.

Yang You, Jing Li, Sashank Reddi, Jonathan Hseu, Sanjiv Kumar, Srinadh Bhojanapalli, Xiaodan Song, James Demmel, Kurt Keutzer, and Cho-Jui Hsieh. 2020. Large batch optimization for deep learning: Training bert in 76 minutes.

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. *Transactions of the Association for Computational Linguistics*.

Xiaohua Zhai, Joan Puigcerver, Alexander Kolesnikov, Pierre Ruyssen, Carlos Riquelme, Mario Lucic, Josip Djolonga, André Susano Pinto, Maxim Neumann, Alexey Dosovitskiy, Lucas Beyer, Olivier Bachem, Michael Tschannen, Marcin Michalski, Olivier Bousquet, Sylvain Gelly, and Neil Houlsby. 2019. The visual task adaptation benchmark. *arXiv*.

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.

A Appendix

A.1 Image-text retrieval

Fine-tuning on Flickr30K and MSCOCO took about 3 and 6 hours, respectively using 32 Cloud TPU V3 cores. All scores of both image-text and text-image retrieval are listed in Table 8 and 9.

A.2 Validation of visual embedding

To validate the visual embedding only, we conduct the ImagenetKNN retrieval task using visual features from our pre-trained model as done in ALIGN. In ImageNet KNN retrieval, we retrieve their nearest neighbors from the training set using pre-trained visual embeddings to find the class of each image in the validation set of ILSVRC-2012. Recall@K metric is obtained using the appearance of the found label of the query image in the top-K retrieved images. We compare our model with ALIGN as shown in Table 10. Because we use image-text pair datasets less than ALIGN by 0.8B pairs, our model shows a score of 67.8 which is still comparable, indicating that our model trained using two alignment task learning ensures powerful visual embedding like ALIGN which is trained with only image-text alignment.

Type	Language	image → text			text → image		
		R@1	R@5	R@10	R@1	R@5	R@10
Zero-shot	en	84.9	97.7	99.5	73.5	92.2	96.0
	de	77.5	96.0	98.4	63.0	86.6	91.9
	fr	70.8	90.6	95.0	72.0	91.2	95.3
	cs	62.8	87.8	93.0	67.2	88.8	93.5
Fine-tune	en	93.5	99.4	99.8	81.8	95.9	98.2
	de	88.1	98.7	99.7	72.1	91.6	95.7
	fr	80.1	96.3	97.9	80.0	95.9	97.8
	cs	77.1	95.0	97.0	77.1	94.5	97.3

Table 8: Image-text retrieval on Flickr30K in multiple languages.

Type	Language	image → text			text → image		
		R@1	R@5	R@10	R@1	R@5	R@10
Zero-shot	en	56.1	80.2	87.5	43.5	69.1	78.4
	ja	35.3	59.9	70.9	25.0	47.3	57.5
	ko	47.1	72.5	82.1	36.8	62.0	72.0
Fine-tune	en	70.2	91.0	95.3	54.9	80.0	87.7
	ja	62.9	87.0	93.2	46.8	73.9	82.9
	ko	69.8	90.8	95.1	53.1	79.1	87.1

Table 9: Image-text retrieval on COCO in multiple languages.

Model	Data	Backbone	ImageNet KNN R@1
ALIGN	1.8B	B7 + BERT-base	69.3
MURAL(reprod.) + ETCL	1.0B + 15M	B7 + BERT-base	67.8

Table 10: Performance on ImagenetKNN retrieval task.

A.3 Ablation study

We investigate the effect of VL pre-training steps with regard to three domain tasks including visual, VL, and language. We select ImageNetKNN and Multi30K(en, de) and STS tasks for three domain respectively.

Performance dependency on VL pre-training steps Table 11 shows the ablation study of VL pre-training steps on three domain tasks. Generally, the longer VL pre-training steps give the stronger performance after ETCL. Interestingly, in the case

Row	VL pre-training steps	ETCL steps	ImageNetKNN R@1	MS-COCO (zero-shot) en(Avg.)	ja(Avg.)	ko(Avg.)	STS Meta avg
1	100K	-	60.5	61.5	29.6	24.0	68.1
2	200K	-	64.0	65.1	34.1	29.1	71.5
3	400K	-	66.0	67.2	37.0	30.1	73.7
4	600K	-	67.5	68.4	40.1	34.1	73.2
5	800K	-	66.9	67.8	40.6	31.5	74.6
6	100K	3K	60.7	61.5	53.2	35.7	78.9
7	200K	3K	64.1	65.2	57.2	42.8	79.9
8	400K	3K	65.6	67.5	57.2	45.7	81.9
9	600K	3K	67.7	69.1	60.3	**51.9**	**81.6**
10	800K	3K	**67.8**	**69.1**	**62.1**	49.3	81.5

Table 11: Performance on visual, VL and language domain tasks with regard to VL pre-training steps.

of Korean MS-COCO, the model after VL pre-training 100K + ETCL has a higher performance by 4.2 than when the model trained after VL pre-training 800K steps. When considering that there is no image and Korean sentence pair in the additional datasets used in ETCL, the result shows that ECTL learns multilingual VL representations very effectively. Likewise, a similar phenomenon is shown in the STS sentence embedding task. This is a good example showing that the visual representation obtained from VL training can be used to improve the performance of text representation.

Model	14 langs	36 langs	82 langs	All langs
LASER	95.3	84.4	75.9	65.5
m-USE	93.0	44.3	38.5	36.6
LaBSE	**95.3**	**95.0**	**87.3**	**83.7**
(Reimers and Gurevych, 2019)	94.8	86.2	75.6	67.0
(Ham and Kim, 2021)	95.4	89.1	79.4	72.9
MURAL(reprod.) + ETCL	88.4	81.5	72.4	63.6

Table 12: Performance on Tatoeba sentence retrieval task. Scores are reported by 100 × accuracy. We follow the grouping '14 langs', '36 langs', and '82 langs' as used in m-USE, XTREME and LASER respectively.

A.4 Zero-shot image classification

Text prompts engineering We use 80 text templates as used in CLIP [7]. For fine-grained datasets, *"a type of food"* are appended to the initial template for adding context information. All multilingual zero-shot ext prompts image classification results in various visual datasets are shown in Table 14.

A.5 Multi-modal zero-shot transfer

To further investigate the effect of multi-modal zero-shot transfer, we calculate the performance of fine-tuned `MURAL(reprod.)` with one language pair from multilingual CC3M on Italian MS-COCO. Interestingly, the performance gains via `ETCL` using the image and regionally closer languages (cs, fr, de) pairs are larger than that of Asian languages (zh, ja), which are regionally far from Italian (See Table 13).

Model	Avg	Gain over baseline
MURAL(reprod.) baseline	40.6	-
MURAL(reprod.) + cs	60.3	19.7
MURAL(reprod.) + fr	60.2	19.6
MURAL(reprod.) + ja	59.7	19.1
MURAL(reprod.) + zh	59.3	18.7
MURAL(reprod.) + de	60.1	19.5

Table 13: Performance on Italian MS-COCO image-text and text-image retrieval tasks. Scores are the average of T2I and I2T R@1,5,10.

A.6 Additional NLP task

Tatoeba For the multilingual sentence retrieval task, we use Tatoeba dataset composed of 1,000 English-aligned sentence pairs for 112 languages (Artetxe and Schwenk, 2019). This task is to find the nearest neighbor for each sentence in another language using the cosine similarity. We conduct an evaluation on three groupings of languages for fair-comparison: the first 14 language groups are selected for m-USE (Yang et al., 2020). The second language group with 36 languages follows the XTREME benchmarks (Hu et al., 2020). The third 82 language group that

[7] https://github.com/openai/CLIP

LASER proposed covers high-resourced to low-resourced languages. Table 12 shows the average accuracy of languages according to different groupings. Our model generally has similar performance compared to LASER, and lower performance than LaBSE. That is because compared to LaBSE and LASER which learns 109 and 93 languages respectively, our model learns 90 languages, and then our model does not support several languages in Tatoeba dataset. For example, Thai language is included in the 14 languages group while the language is not included in the CCMatrix we learned, which leads to a poor performance of 89.2%. The average of 13 languages excluding the language is 94.2%, which is comparable to the top language models. This poor performance from data scarcity also appears for 36 langs groups. The average accuracy of the 36 langs group except for Swahili, Telugu, and Thai is 86.7%, which is better than other models except for LaBSE. Performance degradation due to lack of data in some languages is not relevant to our methodology itself, so it does not impair the novelty of this work. Rather, it is likely that supplementing from other datasets in opus can boost our model. The results for all languages can be found in Table 15, 16.

	af	am	ar	az	be	bg	bn	ca	ceb	cs	cy	da	de
ImageNet	16.5	2.5	18.7	12.9	14.3	21.6	16.3	20.0	21.9	24.6	14.0	30.0	29.3
SUN397	39.0	6.1	43.3	30.9	30.9	46.9	40.2	46.5	33.0	57.2	27.6	56.4	59.5
Food101	51.2	9.6	34.7	47.7	48.9	43.3	47.6	49.6	50.5	62.0	42.6	61.2	66.6
CIFAR100	19.8	6.8	23.7	17.2	15.2	26.0	21.7	23.2	14.9	31.3	10.6	27.2	31.6

	el	en	eo	es	et	eu	fa	fi	fr	fy	ga	gd	gl
ImageNet	18.3	62.6	23.2	25.5	20.2	15.3	15.7	19.7	26.5	24.5	10.4	16.7	28.7
SUN397	45.5	67.2	43.7	48.9	43.5	37.1	40.1	43.3	59.9	37.7	21.0	27.4	58.5
Food101	37.9	74.2	63.8	53.7	40.6	49.4	42.1	47.4	64.6	49.3	45.1	47.2	64.3
CIFAR100	23.8	33.6	25.2	27.6	19.6	19.3	22.7	25.6	32.3	18.0	7.3	14.1	31.1

	ha	he	hl	hr	hu	hy	id	ig	is	it	ja	jv	ka
ImageNet	13.9	14.2	9.8	22.1	20.2	5.8	26.3	12.9	23.4	28.3	17.6	28.2	9.0
SUN397	23.5	35.8	28.9	47.4	46.9	21.3	45.9	21.4	42.8	51.4	52.0	44.0	19.6
Food101	53.5	31.7	31.9	46.4	52.3	33.7	58.3	43.2	54.6	59.4	40.0	51.8	39.4
CIFAR100	13.4	13.7	20.4	23.5	24.7	10.1	26.7	11.1	19.3	25.8	29.8	22.4	9.1

	kk	km	ko	la	lb	lt	lv	mg	mk	ml	mr	ms	my
ImageNet	9.7	5.8	14.1	29.1	20.0	22.8	20.8	20.4	21.8	6.0	8.1	30.3	5.8
SUN397	20.5	12.7	40.0	36.5	36.2	45.1	34.8	40.1	38.3	18.4	23.0	53.6	13.8
Food101	33.4	32.0	33.9	58.9	50.6	52.3	47.8	62.8	51.6	19.8	35.1	60.8	25.1
CIFAR100	14.2	4.8	21.8	20.9	18.6	24.2	19.4	17.5	24.3	12.9	16.9	28.3	4.5

	ne	nl	no	or	pl	pt	ro	ru	sd	si	sk	sl	so
ImageNet	6.2	27.9	28.8	1.1	23.4	26.3	25.9	22.6	2.8	11.0	23.1	28.5	10.2
SUN397	15.4	53.6	54.5	6.3	48.6	50.1	50.5	47.1	8.3	24.7	53.7	50.5	14.1
Food101	15.6	60.9	60.8	10.6	54.3	59.0	53.6	38.3	10.2	29.1	51.1	53.8	38.4
CIFAR100	11.0	26.5	26.5	4.4	25.0	27.3	24.7	26.6	8.6	18.6	28.2	25.7	8.6

	sq	sr	su	sv	sw	ta	tl	tr	tt	uk	ur	uz	vi
ImageNet	21.3	17.0	17.7	25.5	22.5	6.9	36.3	19.7	9.4	18.1	9.8	9.1	23.6
SUN397	39.1	35.7	30.2	53.1	45.2	24.0	53.5	44.1	24.1	46.0	31.5	21.0	49.1
Food101	52.1	17.5	44.7	54.7	65.5	45.7	66.3	50.8	25.0	40.1	46.7	25.1	50.5
CIFAR100	20.6	23.4	18.5	25.9	24.9	10.4	24.8	22.2	13.6	24.5	16.3	12.3	24.8

	xh	yi	yo	zh	zu
ImageNet	12.1	1.8	7.4	21.8	11.5
SUN397	17.3	8.6	13.9	56.0	20.9
Food101	36.9	17.7	35.7	44.7	36.3
CIFAR100	10.3	4.1	5.1	30.9	6.7

Table 14: Zero-shot image classification on ImageNet, SUN397 Food101 and CIFAR100. Scores are 100 × Acc@1.

Model	ar (ara)	bg (bul)	ca (cat)	cs (ces)	cmn	da (dan)	de (deu)	el (ell)	et (est)	fi (fin)	fr (fra)	gl (glg)	he (heb)
LASER	92.0	95.0	95.9	96.5	95.4	96.0	99.0	95.0	96.7	96.3	95.6	95.5	92.2
m-USE	81.0	54.0	66.3	17.8	94.3	25.9	98.2	1.6	8.4	8.2	93.5	82.2	1.8
LaBSE	91.0	95.7	96.5	97.5	96.2	96.4	99.4	96.6	97.7	97.0	96.0	97.2	93.0
MURAL(reprod.) + ETCL	89.8	95.1	96.6	96.6	95.6	96.7	98.8	96.2	96.3	96.4	95.5	95.4	91.5
Model	hi (hin)	hr (hrv)	hu (hun)	hy (hye)	id (ind)	it (ita)	ja (jpn)	ka (kat)	ko (kor)	lt (lit)	lvs	mr (mar)	mk (mkd)
LASER	94.7	97.2	96.0	36.1	94.5	95.3	90.7	35.9	88.9	96.2	95.4	91.5	94.7
m-USE	1.2	23.9	10.2	1.7	93.3	94.3	93.8	2.6	86.0	10.2	11.1	1.8	33.1
LaBSE	97.7	97.8	97.2	95.0	95.3	94.6	96.4	95.9	93.5	97.3	96.8	94.8	94.8
MURAL(reprod.) + ETCL	95.9	97.6	95.7	82.2	94.5	94.3	95.6	66.8	91.7	96.5	93.9	91.8	92.8
Model	mn (mon)	nl (nld)	nb (nob)	pes	pl (pol)	pt (por)	ro (ron)	ru (rus)	sk (slk)	sl (slv)	es (spa)	sq (sqi)	sr (srp)
LASER	8.2	96.3	98.8	93.4	97.8	95.2	97.4	94.6	96.6	95.9	98.0	98.0	95.3
m-USE	16.9	94.0	23.9	12.7	93.7	94.9	30.0	93.7	21.1	20.9	95.4	19.9	27.7
LaBSE	96.6	97.2	98.9	96.0	97.8	95.6	97.8	95.3	97.3	96.7	98.4	97.6	96.2
MURAL(reprod.) + ETCL	18.2	96.5	98.3	95.8	96.7	95.7	96.6	94.6	96.3	96.4	97.5	97.4	94.5
Model	sv (swe)	th (tha)	tr (tur)	uk (ukr)	ur (urd)	vi (vie)	yue	zsm					
LASER	96.6	95.4	97.5	94.5	81.9	96.8	90.0	96.4					
m-USE	18.8	96.0	94.0	51.0	6.4	10.4	84.2	89.1					
LaBSE	96.5	97.1	98.4	95.2	95.3	97.8	92.1	96.9					
MURAL(reprod.) + ETCL	96.4	8.6	98.0	94.9	86.6	97.5	88.2	96.5					

Table 15: Performance on Tatoeba sentence retrieval task for languages. Scores are reported by 100 × accuracy. To make comparisons to other works easily, language abbreviations are expressed using ISO 639-1/639-2/639-3.

Model	af (afr)	am (amh)	ang	arq	arz	ast	awa	az (aze)	be (bel)	bn (ben)	ber	bs (bos)	br (bre)
LASER	89.5	42.0	37.7	39.5	68.9	86.2	36.1	66.0	66.1	89.6	68.2	96.5	15.8
m-USE	63.5	2.1	38.1	28.2	59.6	81.5	2.4	42.2	40.3	0.7	8.3	30.1	10.2
LaBSE	97.4	94.0	64.2	46.2	78.4	90.6	73.2	96.1	96.2	91.3	10.4	96.2	17.3
MURAL(reprod.) + ETCL	92.5	45.2	50.0	42.3	74.2	86.6	44.2	81.6	74.1	89.5	9.6	97.5	12.0
Model	cbk	ceb	ch (cha)	kw (cor)	csb	cy (cym)	dsb	dtp	eo (epo)	eu (eus)	fo (fao)	fy (fry)	gd (gla)
LASER	77.0	15.7	29.2	7.5	43.3	8.6	48.0	7.2	97.2	94.6	71.6	51.7	3.7
m-USE	76.1	13.7	33.6	6.4	37.4	13.1	35.1	8.4	36.8	19.4	18.7	52.3	6.9
LaBSE	82.5	70.9	39.8	12.8	56.1	93.6	69.3	13.3	98.4	95.8	90.6	89.9	88.8
MURAL(reprod.) + ETCL	77.9	22.0	32.9	7.4	45.5	13.7	59.7	11.2	97.4	94.0	72.1	61.9	6.0
Model	ga (gle)	gsw	hsb	io (ido)	ie (ile)	ia (ina)	is (isl)	jv (jav)	kab	kaz	km (khm)	ku (kur)	kzj
LASER	5.2	44.4	54.5	83.7	86.2	95.2	95.6	22.9	58.1	18.6	20.6	17.2	7.2
m-USE	7.7	39.3	33.3	55.5	73.3	86.7	10.3	38.3	3.7	15.3	1.5	21.7	10.2
LaBSE	95.0	52.1	71.2	90.9	87.1	95.8	96.2	84.4	6.2	90.5	83.2	87.1	14.2
MURAL(reprod.) + ETCL	7.2	45.3	65.0	82.7	86.1	93.6	94.9	39.0	4.6	56.0	58.2	21.0	10.6
Model	la (lat)	lfn	ml (mal)	max	mhr	nds	nn (nno)	nov	oc (oci)	orv	pam	pms	swg
LASER	58.5	64.5	96.9	50.9	10.4	82.9	88.3	66.0	61.2	28.1	6.0	49.6	46.0
m-USE	36.7	60.5	1.2	65.0	14.3	57.5	21.2	66.1	42.9	28.3	8.4	48.8	48.7
LaBSE	82.0	71.2	98.9	71.1	19.2	81.2	95.9	78.2	69.9	46.8	13.6	67.0	65.2
MURAL(reprod.) + ETCL	61.2	68.5	97.4	59.9	15.0	68.6	88.8	73.2	63.9	38.9	8.9	56.6	59.8
Model	swh	ta (tam)	tt (tat)	te (tel)	tl (tgl)	tk (tuk)	tzl	ug (uig)	uz (uzb)	war	wuu	xh (xho)	yjd
LASER	57.6	69.4	31.1	79.7	50.6	20.7	44.7	45.2	18.7	13.6	87.7	8.5	5.7
m-USE	13.7	2.8	15.7	2.4	16.2	20.9	46.6	4.0	15.9	15.6	82.2	14.8	1.9
LaBSE	88.6	90.7	97.9	98.3	97.4	80.0	63.0	93.7	86.8	65.3	90.3	91.9	91.0
MURAL(reprod.) + ETCL	71.3	77.2	27.8	4.3	62.3	27.6	52.9	6.0	29.7	24.9	87.3	10.6	8.7

Table 16: Performance on Tatoeba sentence retrieval task for languages. Scores are reported by 100 × accuracy. To make comparisons to other works easily, language abbreviations are expressed using ISO 639-1/639-2/639-3.