Abstract

Vision-and-language pre-training has achieved impressive success in learning multimodal representations between vision and language. To generalize this success to non-English languages, we introduce UC$^2$, the first machine translation-augmented framework for cross-lingual cross-modal representation learning. To tackle the scarcity problem of multilingual captions for image datasets, we first augment existing English-only datasets with other languages via machine translation (MT). Then we extend the standard Masked Language Modeling and Image-Text Matching training objectives to multilingual setting, where alignment between different languages is captured through shared visual context (i.e., using image as pivot). To facilitate the learning of a joint embedding space of images and all languages of interest, we further propose two novel pre-training tasks, namely Masked Region-to-Token Modeling (MRTM) and Visual Translation Language Modeling (VTLM), leveraging MT-enhanced translated data. Evaluation on multilingual image-text retrieval and multilingual visual question answering benchmarks demonstrates that our proposed framework achieves new state of the art on diverse non-English benchmarks while maintaining comparable performance to monolingual pre-trained models on English tasks.

1. Introduction

The world we navigate through is a multimodal and multilingual kaleidoscope. While tremendous success has been realized in multimodal research with the advent of vision-and-language (V+L) pre-training [10, 35, 36, 45, 26], the majority of current literature is biased towards English. Although English-trained V+L models can be finetuned on each target language (given that there is sufficient language-specific data in downstream task), maintaining language-specific models for every language in the world (6,900+) is impossible given insurmountable development and maintenance cost [23]. Naturally, a “Tower of Babel” strategy starts to gain interest in the community, aiming at building one giant model that can handle all languages, notable examples including massively multilingual neural machine translation [1], cross-lingual language model [32], and multilingual multimodal representation learning [18, 24].

Recent release of large-scale multimodal datasets [41] and multilingual corpora (e.g., Wikipedia in 100 languages) has served as a key impetus to accelerate fast advances in V+L pre-training [10, 36, 45, 54] and multilingual language modeling [12, 11, 23], which makes pre-training large-scale multilingual V+L models possible. A pioneering work is M$^3$P [24], which formulates the training process as alternating V+L pre-training between cross-modal monolingual
corpus and mono-modal cross-lingual corpus. It relies on English as the focal point to build a bridge between image and different languages, which inevitably introduces linguistic discrepancy into downstream tasks that rely on direct alignment between image and Non-English language (e.g., Image-to-German retrieval), as shown in Figure 1 (a).

In this paper, we propose a new pre-training framework, UC² (Universal Cross-lingual Cross-modal pre-training), which pivots primarily on images and complementarily on English for multilingual multimodal representation learning (Figure 1 (b)). The major challenge is that pivoting on images requires paired image and aligned multilingual data (e.g., image-English, image-German), while existing V+L datasets only contain image-English pairs. To fill this blank, we propose to augment English-only datasets with other languages via machine translation (MT), and leverage the augmented datasets for pre-training. To the best of our knowledge, this is the first known effort in creating large-scale training datasets with multilingual image captions.

In addition to extending two widely-adopted pre-training tasks (Masked Language Modeling and Image-Text Matching) to a multilingual setting, we further propose two novel pre-training objectives, namely Masked Region-to-Token Language Modeling (MRTM) and Visual Translation Language Modeling (VTLM). MRTM encourages fine-grained alignment between words and image regions, by sharing the embedding space of word tokens and region labels (i.e., object class predictions from an object detector). VTLM is designed to jointly learn cross-lingual cross-modal mapping from parallel textual corpora and paired images. Extensive experiments demonstrate that our proposed UC² framework achieves new state of the art over multiple mainstream benchmarks such as Multi30k [16, 15, 4] and COCO [9, 51, 34] across multilingual image-text retrieval and visual question answering (VQA) tasks.

Our contributions are summarized as follows. (i) We construct a multilingual V+L corpus, and propose the first MT-augmented cross-lingual cross-modal pre-training framework UC², which pivots on both images and English language for joint representation learning. (ii) We propose new pre-training tasks, Masked Region-to-Token Language Modeling and Visual Translation Language Modeling, two effective learning objectives for multilingual multimodal tasks. (iv) We achieve new state of the art on multiple multilingual image-text retrieval and VQA benchmarks, outperforming existing methods.

2. Related Work

Vision-Language Pre-training. There is a growing interest in building generic pre-trained BERT-like [13] models for V+L tasks. Early work such as ViLBERT [36] and LXMERT [45] propose a two-stream architecture that encodes visual and textual input through two separate Transformers, and then fuse the two modalities by a cross-modal Transformer. Later work such as VL-BERT [44], Unicoder-VL[33] and UNITER [10] introduce a single-stream architecture that uses one Transformer to encode concatenated input from both modalities simultaneously. Later, Unified VLP [54] applies to both understanding and generation tasks. Further improvements are proposed on using different input features [35, 26] and multi-task learning [37].
Multimodal Multilingual Learning. Existing studies arching over multilingual and multimodal aspects mainly focus on two tasks: cross-modal retrieval and multimodal machine translation (MT). [38, 7] introduces a multimodal multilingual approach by aligning images and captions in different languages to English captions. Unlike previous work using languages as a pivoting point, [18] learns a shared embedding space that forces representations of different languages towards the pivot image representation. Later work focuses on scaling to more languages via character-based word-embedding [49] or shared language-acoustic embedding [28]. SMALR [5] proposes a scalable multilingual model to learn visually aligned word embeddings, for better balance between multilingual capacity and task performance.

Multimodal MT exploits visual information to improve language translations. Earlier work introduces vision to an LSTM-based neural MT model via attention to visual context [8, 21], or fusion [6], or multi-task learning [17, 55]. Lately, Transformer-based [46] models are proposed [2, 50]. There is also an growing interest in unsupervised multimodal MT [25, 43], where translation between monolingual corpus is augmented via pivoting on image.

While successful in individual tasks, these models are usually trained on small amount of data, which limits its extension to other tasks or languages. To learn task-agnostic universal representations across vision and multilingual text, M^3P [23] introduces the first pre-training framework that alternatively optimizes the model on multimodal monolingual corpus and mono-modal multilingual corpus. While M^3P achieves better performance compared to task-specific methods, the alignment between vision and Non-English languages is hard to capture, as the model is learned via using English as the anchor point. To strengthen the alignment between vision and all languages, we propose to pre-train a unified architecture where sentences in different languages are grounded on shared visual context.

### 3. Cross-Lingual Cross-Modal Pre-training

In this section, we start with introducing our machine translation augmented dataset that enables large-scale cross-lingual pre-training. We then go over the proposed UC^2 model and our designed pre-training objectives for universal representation learning across vision and languages.

3.1. Machine Translation Augmented Dataset

Our multilingual image-text paired data is collected via augmenting the captions from the Conceptual Captions dataset [41] with a set of machine translated captions in other languages \( L = \{l_1, l_2, \ldots, l_n\} \). Specifically, we translate the original English captions into five different languages (German, French, Czech, Japanese, and Chinese), which covers languages required for all the downstream tasks studied in this work. Note that with recent advances on machine translation for low-resource languages, we can further expand the dataset to more languages, which we leave for future work. With this data augmentation, we obtained 3.3 million images, each paired with captions in six languages, as the process shown in Figure 2 (a). This one-to-many mapping greatly facilitates the learning of alignment between visual content and semantics from each language through image as a shared anchor. By introducing translated data into model pre-training, our method yields significant improvement over the baseline with MT tools applied only on downstream tasks. Next, we elaborate how to leverage these data for cross-lingual cross-modal pre-training.

3.2. Model Overview

UC^2 extends monolingual language encoder of V+L frameworks, such as UNITER [10], to cross-lingual encoder [11], as shown in Figure 2 (b). The visual feature is extracted from an image encoder and the language feature is obtained from a general cross-lingual language encoder. The multimodal features are then combined into a sequence and fed to a multi-layer Transformer to produce contextualized cross-modal and cross-lingual representations.

Image Encoder. Given an input image, we first obtain a sequence of image region features \( \mathbf{v} = \{v_1, v_2, \cdots, v_m\} \) with Faster R-CNN [39]. For each region, we also extract location features via a 7-dimensional vector: \( \mathbf{p} = [x_1, y_1, x_2, y_2, w, h, w \times h] \), which denotes the normalized top left coordinates, bottom right coordinates, width, height, and the area of the detected region box. The region feature and location feature are fed through separate fully-connected (FC) layers to be projected into the same dimension as the text embedding space, followed by a layer-normalization (LN) layer. The final representation of the region feature is then obtained via summing up the projected region feature and location feature.

Cross-lingual Language Encoder. We follow XLM-R [11] to tokenize an input sentence \( T^{i_l} \) in language \( l_i \) to BPE tokens \( t^{i_l} = \{t^{i_l}_1, t^{i_l}_2, \cdots, t^{i_l}_n\} \) using Sentence Piece model [31]. We then project each token to its embedding based on the XLM-R vocabulary and word embeddings. The final representation of each token is obtained via summing up its word embedding, segment embedding, and position embedding as in XLM-R, followed by another Layer Normalization.

3.3. Pre-training Tasks

For model training, we employ four pre-training objectives to train on large multilingual image-text paired data:
3.3.1 General Tasks

Following previous V+L pre-training work [10, 33, 36, 44], we consider Masked Language Modeling and Image-Text Matching as two of our pre-training tasks.

Masked Language Modeling (MLM). Given a set of image regions \( v = \{v_1, v_2, \ldots, v_m\} \) and its associated caption words \( w^l = \{w^l_1, \ldots, w^l_T\} \) in language \( l_i \in L \), and mask indices as \( m \in \mathbb{N}^M \), we randomly mask a word \( w^l_m \) with the probability of 15% and replace the masked word with a special token [mask]. The objective is to predict the masked word \( w^l_m \) based on the surrounding words \( w^l_{nm} \) and all image regions \( v \), by minimizing the negative log-likelihood:

\[
L_{MLM}(\theta) = -\mathbb{E}_{(w^l_i, v) \sim D} \log P_{\theta}(w^l_i | w^l_{\sim m}, v),
\]

where \( \theta \) is the learnable parameters. Each pair \((w^l_i, v)\) is sampled from the whole training set \( D \). The caption for each language is sampled with even probability \( p = 1/|L| \).

Image-Text Matching (ITM). ITM has been widely used in vision-and-language pre-training [10, 33, 36, 44] to learn instance-level alignment between image and sentence. The output of the special token [cls] is fed through a FC layer and a sigmoid function to predict a score \( s_\theta(v, w^l) \) between 0 and 1, which predicts whether the input image \( v \) and the text input \( w^l \) are semantically matched. During training, we sample positive and negative pairs from the dataset \( D \) with equal probability at each step. The negative image-text pair is created by replacing the image or text in a matched pair with a randomly-selected distractor from the same mini-batch. The objective is optimized with binary cross-entropy loss:

\[
L_{ITM}(\theta) = -\mathbb{E}_{(w^l_i, v) \sim D}[y \log s_\theta(w^l_i, v) \\
+ (1-y) \log (1 - s_\theta(w^l_i, v))],
\]

where \( y \in \{0, 1\} \) indicates whether the input image-text pair is a positive or negative sample. The deployment of MLM and ITM serves as our base model. Next, we introduce two novel objectives to further enhance cross-lingual cross-modal representation learning.

3.3.2 Masked Region-to-Token Modeling

Now that we have a learning objective for language (MLM), how about the vision counterpart? In existing VLP models, Masked Region Modeling (MRM) serves this purpose by predicting the top-1 or soft object label associated with the masked image region. The de facto approach to harvesting object labels is using the predictions from an off-the-shelf object detector \( \text{e.g., Faster R-CNN} [39] \). However, there are two limitations with this approach. First, the association between object labels from image and word tokens from text is not well utilized. While salient objects detected in the image are usually mentioned in the paired description, MRM misses this connection as it directly predicts masked image region to an index between 0 and 1600. Second, the visual embedding extracted from an object detector can differ significantly from pre-trained word embedding due to different embedding space. Existing methods merely rely on weak supervision from pre-training objectives to close the gap between these two disparate embedding spaces. We argue that a well-aligned embedding space is indispensable for our problem, given its complex multilingual multimodal nature. Therefore, we propose to explicitly learn the correspondence between region and word tokens and tackle the aforementioned issues with two strategies.

Masked Region-to-Token Modeling (MRTM). This new objective aims to classify each masked region to its “pseudo” object label \( \text{e.g., dog, cat, provided by a pre-trained object detector) \), which is the (sub-word) token in our word vocabulary that associates with the original object label. Compared to the MRM objective from previous work [36, 33, 10], MRTM leverages additional semantic association between object labels and captions to capture semantic alignment between vision and language. More formally, given an image region \( v_i \in v \), we set its probability for being masked out as 15% (as in [13]). For each masked region, the region feature vector is either replaced by a zero-initialized vector \( v_m \) (90% probability) or remains the same (10%). Then we predict the associated “pseudo” object label \( c^l_{vm} \) on the masked region based on the observation of surrounding image regions \( v\sim m \) and the paired caption \( w^l \) in language \( l_i \), by minimizing the negative log-likelihood:

\[
L_{MRTM}(\theta) = -\mathbb{E}_{(w^l_i, v) \sim D} \log P_{\theta}(c^l_{vm} | w^l_i, v\sim m)
\]

Early Adaptation (EA). To address the second limitation and facilitate learning of a joint embedding space between vision and language, we warm up the image encoder to make sure the output visual embedding shares the same embedding space as word embeddings. Specifically, each image region is projected to an image region feature \( v_i \in \mathbb{R}^p \) through the image encoder, with the same dimension as the

\footnote{When multiple tokens are associated to one label word, we randomly select one to decode during pre-training}
word embedding vector. We then extract the word embedding vectors from XLM-R that correspond to the $k$ object categories $c = \{c_1, c_2, \ldots, c_k\}$ defined by the object detector. We compute the cosine similarity between the projected image feature $v_i$ with the $k$ word embedding vectors followed by a softmax function, resulting in a normalized distribution $h_{\theta_l}(v_i) \in \mathbb{R}^k$ that indicates the prediction on what semantics are mapped in the region. We then maximize the similarity between this predicted distribution and the “GT” object probability distribution from the object detector output $g(v_i) \in \mathbb{R}^k$, by minimizing their KL divergence:

$$
L_{EA}(\theta_I) = D_{KL}(g(v_i)||h_{\theta_l}(v_i)),
$$

where $\theta_I$ is the learnable parameters of the image Encoder.

Note that a recent work named OSCAR [35] has made a similar effort to close the visual-textual embedding gap by inserting object tags into the input sequence. Compared to OSCAR [35], our method has two advantages. First, it does not rely on object tags for downstream tasks, which might not be applicable for image domains that cannot be well covered by the object categories from the pre-trained detector. Second, by forcing image representation to be similar to language representation with EA, our pre-training model can better leverage the initialized weights from language-only pre-trained model to adapt to image modality.

### 3.3.3 Visual Translation Language Modeling

All the objectives mentioned so far operate on image and monolingual input, without considering cross-lingual objectives. The correspondence between languages is vital for cross-lingual generalization, clearly observed from existing work on language understanding [11]. Our proposed methods so far unexceptionally learn cross-lingual correspondence indirectly through the image focal point, which might not be sufficient. We hence propose visual translation language modeling (VTLM), which directly and jointly learns the alignment between visual context and text in different languages.

In VTLM, given an image $v$ and a pair of captions $(u_i^{l1}, u_i^{l2})$ in two different languages, the goal is to predict masked caption tokens from both languages. One of the two languages is always English, as English captions in our pre-training data are directly from [41], while captions in other languages are translated by MT, therefore less reliable. Under this bilingual framework, model input size only grows linearly with more languages.

Besides, as our model is initialized with the weights of a powerful pre-trained multilingual model, it has already learned a good alignment between different linguistic words to some extent. Applying random masking strategy in VTLM is sub-optimal, as the model can make a correct prediction by simply translating words from one language to another, without taking into account the visual information from image. To encourage the model to fully consider visual context, we introduce a strategy called co-masking, where we simultaneously mask out tokens with similar semantic meanings from paired captions to prevent easy translations.

There are a few steps in co-masking. First, we apply Fast Align [14] to learn the word alignment between two different languages $(l_i, l_j)$ from the noisy parallel corpus that was created using machine translation. Then, during the pre-training stage, we follow the same strategy as in MLM to randomly mask a token $w^{l_i}_{m}$ from the caption of one language. For the paired caption in the other language $l_j$, we mask the aligned word tokens $w^{l_j}_{m}$ that are predicted from Fast Align. [14] The final objective is again to predict masked tokens from both languages by minimizing the negative log-likelihood:

$$
L_{VTLM}(\theta) = -\mathbb{E}_{(w^{l1}, w^{l2}, v) \sim D} \log P_{\theta}(w^{l1}_{m}, w^{l2}_{k} | w^{l1}_{m}, w^{l2}_{k}, v)
$$

### 4. Experiments

In this section, we provide detailed experiments to evaluate our proposed UC² model over multilingual image-text retrieval and multilingual VQA tasks.

**Multilingual Image-Text Retrieval** In the retrieval task, the model retrieves an image from a set of candidates given a caption in a certain language, or vice versa. We consider two datasets: Multi30K [16, 15, 4] and MSCOCO [9, 51, 34]. Multi30K is built upon Flickr30K [52], where English captions are manually translated to German, French, and Czech. It contains 31K images (each paired with 5 captions in English and German, 1 caption in French and Czech). Following Flickr30K [52], we split the data into 29K/1K/1K images for train/val/test.

MSCOCO [9] consists of 123K images, with 5 English captions per image. STAIR [51] extends MSCOCO dataset by collecting 820K Japanese captions for 165K COCO images. Similarly, Li et al. [34] collect Chinese captions for 20K COCO images with roughly 1 caption per image. We use the train/dev/test splits for English and Japanese defined in [27], and present results on the 1K test set. For MSCOCO Chinese, we follow the original split as in [34]. We compute Recall@K (recall of top K candidates) for both image-to-text retrieval and text-to-image retrieval with $K = 1, 5, 10$. The average of all these 6 evaluation scores, Average Recall (AR)[24], is used as the final evaluation metric.

**Multilingual Visual Question Answering (VQA)** In Multilingual VQA, given an image and a question in a certain language, the model predicts an answer based on the visual context in the image. We evaluate our model on two
datasets: VQA v2.0 [20] and Japanese Visual Genome (VG) VQA [42]. VQA v2.0 is a widely used benchmark for English VQA task. We follow the official partition to split the dataset and report results on Test-Dev set through the official evaluation server. Following [10], our training is augmented by running on both the training and validation split of VQA v2.0 as well as the VQA from Visual Genome [30]. Visual Genome VQA Japanese [42] expands the VG English VQA dataset [30] by collecting 793K Japanese question answering pairs on 99K images from VG. We use the train/test split in the original VG VQA to split the data into 61K/30K training/test images.

We formulate VQA as a multi-label classification problem, where the model predicts answer from the candidate pool. VQA score [20] is used to compare model predictions against 10 human-annotated answers in VQA v2.0. On Visual Genome VQA Japanese, which only has one ground-truth answer to each question, we use accuracy and BLEU score as the evaluation metrics.

### Implementation Details

UC² consists of 12 layers of transformer blocks, where each block has 768 hidden units and 12 self-attention heads. Except for the image encoder, the model is initialized with XLM-R [11]. We run continuous pre-training with MLM, ITM, MRTM and VTLM objectives. We use Adam optimizer [29] with a linear warm-up for the first 5% of training, and set the learning rate to $4e^{-4}$. We use Horovod and NCCL for multi-node communications and apply gradient accumulation (every 3 steps) to reduce multi-GPU communication overheads. The batch-size for pre-training is set as 1024 and the dropout rate is 0.1. Pre-training experiments are conducted on 8 Nvidia V100 GPUs for 30 epochs, which takes 4 days to converge.

### 4.1. Experimental Results

We first compare UC² to various SOTA with or without pre-training on the two downstream tasks. Then, We conduct ablation experiments to study the effectiveness of MRTM and VTLM, as well as the impact of image pivoting. Finally, we visualize the alignments between visual context and cross-lingual text context learned by our pre-trained UC² model.

#### 4.1.1 Evaluation on Multilingual Retrieval

We compare UC² with state-of-the-art methods on image retrieval and text retrieval in two different settings:

- **English-only Fine-tune**: We fine-tune the pre-trained model on just the English training data.
- **Single-Language Fine-tune**: We fine-tune the pre-trained model on training data for each target language.
- **All-Language Fine-tune**: We fine-tune the pre-trained model on merged training data of all languages.

Besides reporting AR on each language, we also compute the Meta-Ave (average of AR across all languages over two datasets) to reflect the overall performance in this task. Given that we have access to pre-trained machine translation models, we also introduce a strong translate-test baseline UNITERcc based on [10], which is pre-trained on Conceptual Conception English data and finetuned on English.

| Method          | Flickr30K | MSCOCO |
|-----------------|-----------|--------|
| EN DE FR CS EN ZH JA Meta-Ave |           |        |
| SOTA without pre-training |           |        |
| EmbN[47]        | 72.0      | 60.3   |
| PAR.EmbN [19]   | 69.0      | 62.6   |
| S-LIWE [49]     | 76.3      | 72.1   |
| MULE [28]       | 70.3      | 64.1   |
| SMALR [5]       | 74.5      | 69.8   |
| English-only Fine-tune |           |        |
| M³P [24]        | 87.4      | 85.8   |
| UC²             | 87.2      | 74.9   |
| Translate-Test  |           |        |
| UNITERcc [10]   | 87.7      | 81.2   |
| Single-Language Fine-tune |           |        |
| M³P [24]        | 87.4      | 82.1   |
| UC²             | 87.2      | 83.8   |
| All-Language Fine-tune |           |        |
| M³P [24]        | 87.7      | 82.7   |
| UC²             | 88.2      | 84.5   |

Table 1. Evaluation results on image-text retrieval over Flickr30K and MSCOCO datasets across different languages. We highlight the MSCOCO results for MULE and SMALR in blue as they are using different dev/test splits of MSCOCO compared to other models.
training data in the downstream tasks. By translating the test data from other languages to English, \textsc{UNITER} can be directly applied for text/image retrieval. Results are summarized in Table 1.

Our model on the all-language setting achieves a significant improvement over all task-specific methods without pre-training, showing the effectiveness of cross-lingual cross-modal pre-training in learning universal representation across vision and different languages. Our model also demonstrates a superior transferability. When finetuned on English dataset only, we observe an absolute gain of 17.3% on Meta-Ave across different languages over M3P via better transition of the learned knowledge from English to other languages. Compared to the best non-pre-trained models trained on data in each language, our cross-lingual model under the English-only fine-tune setting is still 5% better. We suspect the improvement comes from the in-domain pre-training objective: we use image as the grounding media in ITM to learn cross-modal mapping from one language to another. With strong transfer capability, our model could potentially generalize the learned knowledge from a high-resource language to downstream tasks in low-resource languages.

When we finetune UC² model on all-language data, our model still demonstrates a consistent advantage over M³P on the majority of languages, with 3.4% improvement on Meta-Ave. Our best model is also better than the strong translate-test baseline \textsc{UNITER} on all languages except English in MSCOCO. The slightly worse performance on COCO English is potentially due to lack of pre-training in English data, given that our pre-training time is evenly split to multiple languages. However, this does not overshadow the fact that we achieve overall better performance across all languages. Thanks to the cross-lingual pre-training and finetuning, our model can leverage the complementary information captured in different languages to improve the performance on each language.

### 4.1.2 Evaluation on Multilingual VQA

For multilingual VQA, our pre-trained model is finetuned and evaluated on the target language for each dataset. Unlike image-text retrieval where the same output layer is shared across different languages, multilingual VQA has different classes of answers for each language, which makes joint training across different languages impossible. We compare our model with state-of-the-art methods without pre-training as well as V+L pre-training methods that use the same pre-training corpus. When evaluating the translate-test baseline \textsc{UNITER} on VG VQA Japanese dataset, we first finetune it on VQA v2.0 [20] with English answer candidates translated from VQA VG Japanese to ensure the same reference is used during evaluation as in UC². We then use machine translation model to translate the test dataset of VG VQA Japanese to English, and evaluate the finetuned translate-test model using classification accuracy and BLEU. Results are summarized in Table 2.

On VQA v2.0, our model achieves significant improvement over SOTA task-specific method, and also outperforms existing monolingual models pre-trained on Conceptual Conception [36, 44, 10] by an obvious margin. On VG VQA Japanese, we finetune our model with a different data split from the original baseline method PCATT proposed in VG VQA Japanese, where we have much less training data than their split (ours: 61K images vs. PCATT: 91K images). Even under this disadvantage in an unfair comparison, our pre-trained model still achieves more than 10% improvement on both accuracy and BLEU over baselines. Although achieving better performance compared to the task-specific method, the translate test baseline (\textsc{UNITER}/cc) performs much worse than UC² on the translated VQA VG Japanese dataset. Despite strong performance on the VQA English dataset, the noisiness from the machine translated language would lead to unavoidable degradation especially for tasks like VQA that requires fine-grained level understanding and interpretation on multi-modal context. Hence, building unified cross-lingual cross-modal pre-training model like UC² is a better solution to directly work on tasks in target languages than a translate-test method.

#### 4.1.3 Ablation Study

**Effect of Training Objectives** To validate the effectiveness of the proposed pre-training objective MRTM and VTLM, we conduct ablation study to verify their contributions to the model performance. We gradually remove the two proposed training objectives and evaluate these ablated models on our two downstream tasks. When finetuning the pre-trained model on the image-text retrieval task, we follow the best experimental setting to train the model on all language data. On VQA task, the model is directly fine-tuned on the target language data.

From Table 3, we observe that MRTM has led to significant performance boost on multilingual VQA tasks over the
two languages while gaining some incremental improvement on image-text retrieval tasks. VQA requires more fine-grained understanding about connections between language and visual context, therefore benefits more from the cross-modal local alignment captured by MRTM. When introduce VTLM to the pre-training of UC2, we observe similar improvement on image-text retrieval task, but the improvement on VQA VG Japanese is relatively incremental. We suspect the limited help is mainly due to the language gap between English and Japanese captions. Hence it is hard to capture the good alignment between English and Japanese via VTLM.

**Effect of Pivoting on Image** To validate the effectiveness of image pivoting, we conduct a controlled experiment where the model variant only pivots on English. In this setting, we train UC2 with all the pre-training objectives on English Conceptual Caption data, except for VTLM which involves image as one of the pivoting points. To capture the alignment between English and other languages, we train UC2 on pairs of captions in two different languages with one language fixed as English. The training objective is translated language modeling adopted from XLM [32]. From Table 4, we can see that UC2 pre-trained by pivoting on image achieves overall better performance in multilingual image-text retrieval task. The advantage is particularly sound when the target language has limited training data. This indicates that the cross-lingual cross-modal representation learned by pivoting on images imbues stronger cross-modal mapping transfer across different languages.

**Visualization** To visualize the cross-lingual cross-modal alignment learned by UC2, we provide examples of text-to-image attention from salient words in multilingual captions to the shared image context. As shown in Figure 3, words from different languages that share the same semantic meaning can attend to similar corresponding regions in the image. This shows that while our model can effectively capture cross-modal alignments between regions and words, it also connects different languages by grounding them to similar image regions.

## 5. Conclusion

We present the first MT-augmented pre-training model UC2 that pivots primarily on images and complementary on English to learn cross-lingual cross-modal representation from large scale of multilingual image-to-text pairs. We propose two new pre-training tasks that facilitate our model to capture better alignment between vision and different languages. Our model achieves the new state-of-art performance on two mainstream multilingual V+L tasks and demonstrate strong cross-lingual transfer capability. For future work, we will continue exploring this topic and expanding the framework to include more families of languages. As more benchmarks [48, 40, 3] on multilingual video-text pairs become available, we are interested in enhancing the grounding between vision and language by leveraging the temporal information from videos.
References

[1] Roei Aharoni, Melvin Johnson, and Orhan Firat. Massively multilingual neural machine translation. In North American Chapter of the Association for Computational Linguistics, 2019.

[2] Hasan Sait Arslan, Mark Fishel, and Gholamreza Anbarjafari. Doubly attentive transformer machine translation. CoRR, abs/1807.11605, 2018. 1, 3

[3] AmirAli Bagher Zadeh, Yansheng Cao, Simon Hessner, Paul Pu Liang, Soujanya Poria, and Louis-Philippe Morency. CMU-MOSEAS: A multimodal language dataset for Spanish, Portuguese, German and French. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1801–1812, Online, Nov. 2020. Association for Computational Linguistics. 8

[4] Loïc Barrault, Fethi Bougares, Lucia Specia, Chiraag Lala, Desmond Elliott, and Stella Frank. 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, Belgium, Brussels, Oct. 2018. Association for Computational Linguistics. 2, 5, 12

[5] Andrea Burns, Donghyun Kim, Derry Tanti Wijaya, Kate Saenko, and Bryan A. Plummer. Learning to scale multilingual representations for vision-language tasks, 2020. 1, 3, 6

[6] Iacer Calixto and Qun Liu. Incorporating global visual features into attention-based neural machine translation. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 992–1003, Copenhagen, Denmark, Sept. 2017. Association for Computational Linguistics. 1, 3

[7] Iacer Calixto and Qun Liu. Sentence-level multilingual multi-modal embedding for natural language processing. In Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017, pages 139–148, Varna, Bulgaria, Sept. 2017. INCOMA Ltd. 3

[8] Iacer Calixto, Qun Liu, and Nick Campbell. Doubly-attentive decoder for multi-modal neural machine translation. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1913–1924, Vancouver, Canada, July 2017. Association for Computational Linguistics. 3

[9] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO captions: Data collection and evaluation server. CoRR, abs/1504.00325, 2015. 2, 5, 12

[10] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. {UNITER}:: Learning {un}iversal image-{te}xt representations, 2020. 1, 2, 3, 4, 6, 7

[11] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. arXiv preprint arXiv:1911.02116, 2019. 1, 3, 5, 6

[12] Alexis Conneau and Guillaume Lample. Cross-lingual language model pretraining. In Advances in Neural Information Processing Systems, pages 7059–7069, 2019.

[13] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. 2, 4

[14] Chris Dyer, Victor Chahuneau, and Noah A. Smith. A simple, fast, and effective reparameterization of IBM model 2. In Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 644–648, Atlanta, Georgia, June 2013. Association for Computational Linguistics. 5

[15] Desmond Elliott, Stella Frank, Loïc Barrault, Fethi Bougares, and Lucia Specia. Findings of the second shared task on multimodal machine translation and multilingual image description. In Proceedings of the Second Conference on Machine Translation, pages 215–233, Copenhagen, Denmark, Sept. 2017. Association for Computational Linguistics. 2, 5, 12

[16] Desmond Elliott, Stella Frank, Khalil Sima’an, and Lucia Specia. Multi30K: Multilingual English-German image descriptions. In Proceedings of the 5th Workshop on Vision and Language, pages 70–74, Berlin, Germany, Aug. 2016. Association for Computational Linguistics. 2, 5, 12

[17] Desmond Elliott and Ákos Kádár. Imagination improves multimodal translation. In Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 130–141, Taipei, Taiwan, Nov. 2017. Asian Federation of Natural Language Processing. 3

[18] Spandana Gella, Rico Sennrich, Frank Keller, and Mirella Lapata. Image pivoting for learning multilingual multimodal representations. In Empirical Methods in Natural Language Processing, 2017. 1, 3

[19] Spandana Gella, Rico Sennrich, Frank Keller, and Mirella Lapata. Image pivoting for learning multilingual multimodal representations. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 2839–2845, Copenhagen, Denmark, Sept. 2017. Association for Computational Linguistics. 1, 6

[20] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. In Conference on Computer Vision and Pattern Recognition (CVPR), 2017. 6, 7, 12

[21] Jindřich Heřel and Jindřich Libovický. CUNI system for the WMT17 multimodal translation task. In Proceedings of the Second Conference on Machine Translation, pages 450–457, Copenhagen, Denmark, Sept. 2017. Association for Computational Linguistics. 3

[22] John Hewitt, Daphne Ippolito, Brendan Callahan, Reno Kriz, Derry Tanti Wijaya, and Chris Callison-Burch. Learning
[47] Liwei Wang, Yin Li, and Svetlana Lazebnik. Learning two-branch neural networks for image-text matching tasks. *CoRR*, abs/1704.03470, 2017.

[48] Xin Wang, Jiawei Wu, Junkun Chen, Lei Li, Yuan-Fang Wang, and William Yang Wang. Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.

[49] Jonatas Wehrmann, Douglas M. Souza, Mauricio A. Lopes, and Rodrigo C. Barros. Language-agnostic visual-semantic embeddings. In *International Conference on Computer Vision*, 2019.

[50] Shaowei Yao and Xiaojun Wan. Multimodal transformer for multimodal machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4346–4350, Online, July 2020. Association for Computational Linguistics.

[51] Yuya Yoshikawa, Yutaro Shigeto, and Akikazu Takeuchi. STAIR captions: Constructing a large-scale Japanese image caption dataset. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 417–421, Vancouver, Canada, July 2017. Association for Computational Linguistics.

[52] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014.

[53] Zhou Yu, Jun Yu, Yuhao Cui, Dacheng Tao, and Qi Tian. Deep modular co-attention networks for visual question answering. *CoRR*, abs/1906.10770, 2019.

[54] Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason J. Corso, and Jianfeng Gao. Unified vision-language pre-training for image captioning and vqa. In *AAAI*, 2020.

[55] Mingyang Zhou, Runxiang Cheng, Yong Jae Lee, and Zhou Yu. A visual attention grounding neural model for multimodal machine translation. In *Empirical Methods in Natural Language Processing*, 2018.
A. Appendix

In this supplementary materials, we present the implementation details for downstream task finetuning (Section A.1) and more ablation study on the proposed pre-training objectives (Section A.2).

A.1. Downstream Tasks Details

Multilingual Image-Text Retrieval During fine-tuning, we train and evaluate the pre-trained UC$^2$ on Multi30K [16, 15, 4] and MSCOCO [9, 51, 34]. When we fine-tune UC$^2$ on both datasets, we use batch size of 40 and sample 2 negative image-text pairs for each sampled positive image-text pair. The pre-trained model is optimized by the Adam Optimizer with the learning rate set to $1 \times 10^{-4}$ and a linear warm-up for the first 10% of fine-tuning. For Cross-Lingual zero-shot setting, the pre-trained UC$^2$ is fine-tuned on English-only training data for 30K steps. For All-Language setting, we train UC$^2$ on all the training data in all languages for 50K steps. The finetuning is run on 8 Nvidia V100 GPUs.

Multilingual VQA When we fine-tune the pre-trained model on VQA, the output layer is formatted to output a probability distribution over a set of predefined answers. To train the pre-trained model on VQA, we apply a binary cross-entropy loss and optimize the objective with Adam optimizer. On VQA v2.0 [20], we set batch size to 10240 and the learning rate as $2 \times 10^{-5}$, and the model is trained for 7K steps. On VQA VG Japanese [42], the model is trained for 5K steps with the batch of 5120 and the learning rate of $8 \times 10^{-5}$. The weight decay for the fine-tuning on both datasets is set to 0.01. The fine-tuning for VQA is run on 4 Nvidia V100 GPUs.

| Pre-training tasks | ITR Meta-Ave | VQA EN | VQA JA |
|--------------------|--------------|--------|--------|
| ITM+MLM+MRC        | 85.1         | 70.60  | 33.4   |
| ITM+MLM+MRTM       | 85.3         | 71.45  | 34.1   |

Table 5. Direct ablation on comparison between the proposed MRTM and the MRC. The presentation of the result is simplified to only include the Meta-Average for the multilingual image-text retrieval over both Multi30K and MSCOCO, the accuracy on VQA v2.0 test-dev split (referred as VQA EN), and the accuracy on VQA VG Japanese (referred as VQA JA).

A.2. Ablation Studies

In this section, we provide more ablation studies on our proposed pre-training objectives that could not fit in the main paper due to space limit. First, we provide evidence to show our proposed MRTM is better than the traditional masked region modeling task MRC for multilingual multimodal pre-training. Second, we study the impact of number of languages included in pre-training data. Last, we explore the effect of using different pivoting languages in our proposed VTLM task.

MRTM vs MRC For this ablation, we pre-train UC$^2$ with ITM, MLM and MRC and compare the results to the pre-trained UC$^2$ optimized with ITM, MLM and MRTM. The results is summarized in Table. 5. As shown in table 5, compare the pretrained UC$^2$ that employs the traditional task MRC and the one that employs our proposed MRTM, we can see that the performance on the image-text retrieval task are similar, but MRTM leads to marginal improvement on the multilingual VQA tasks. This observation is consistent with our hypothesis that the proposed MRTM augments the local alignment between image regions and the words in different languages which benefits downstream tasks that rely on region-level recognition and reasoning.

Effect of Pre-training languages As we use machine translation models to expand the pre-training corpus, theoretically, we can have as many languages as needed. We conduct further experiments to verify the impact of number of languages included in pre-training data. We create three variants of pre-training corpus, where the number of languages are 2, 3, and 6, respectively. Every corpus contains English and German. We add Chinese to construct the corpus with 3 languages, and the corpus with 6 languages contains all the languages used to pre-train our full model. The pre-trained models are evaluated on image-text retrieval task in English and German, by finetuning on target language.

Figure 4 shows that when the number of pre-training languages increases, the performance on image-text retrieval on different languages (English and German) slightly improves. This result demonstrates that cross-lingual cross-modal pre-training can effectively leverage different lan-

---

5For fair comparison, we constraint the training time to be the same with different pre-training corpus.
guages to learn stronger vision-to-monolingual-sentence alignment. Meanwhile, as we maintain the same pre-training epochs for all three experiments, we also observe that the benefit of multilingual V+L pre-training is compensating for the reduced training time allocated to each language. Although more comprehensive analysis in future study can help us better understand the trade-off between language capacity and performance on downstream tasks, our observation to some extent still suggests that our model is scalable to pre-training on a large corpus with many languages within a reasonable time frame.

**Effect of Pivoting Language in VTLM** We also conducted a controlled experiment to learn the effect of different pivoting languages in VTLM for the multi-lingual multi-modal pre-training. In this controlled experiment, we pre-train UC2 with all the objectives but change the pivoting language in VTLM from English to Chinese. When we evaluate the pre-trained model on the multilingual image-text retrieval task, the meta-ave score for the pre-trained model with VTLM pivoted on Chinese is dropped from 86.2 to 85.5. This to some extent suggest that English is a more optimal pivoting language to learn the cross-lingual cross-modal shared representation space. Another potential reason for the limited performance is due to the noiseness in the pre-trained Chinese captions gained via automatic machine translation. To gain more solid conclusion to determine the optimal pivoting language, more comprehensive experiments need to be conducted in the future work.