UC^2: Universal Cross-lingual Cross-modal Vision-and-Language Pre-training

Mingyang Zhou[1], Luowei Zhou[2], Shuohang Wang[2], Yu Cheng[2], Linjie Li[2], Zhou Yu[1], Jingjing Liu[2]

[1]University of California, Davis
[2]Microsoft Dynamics 365 AI Research

{minzhou, joyu}@ucdavis.edu
{luozhou, shuowa, yu.cheng, lindsey.li, jingjl}@microsoft.com

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 main-
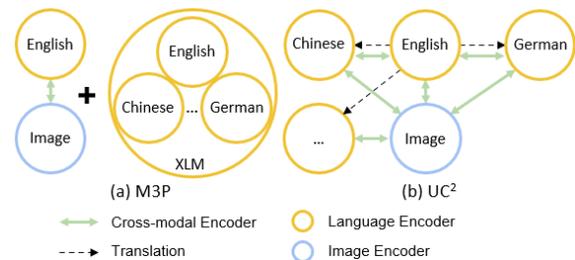
Figure 1. A topology comparison between existing work (M3P) and our proposed UC^2. M3P combines two types of pre-training tasks and the cross-modal Transformer works only on images and English captions. Our UC^2 builds a cross-lingual cross-modal Transformer over images and all the other languages.

tenance 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].

Early works on cross-lingual multimodal tasks mainly focus on machine translation [22, 55, 6, 50, 2] and image-text retrieval [18, 28, 5, 19, 49]. The goal is to construct a common embedding space for vision and cross-lingual inputs, and draw visual concepts from images and similar semantics from languages close together in the feature space. However, due to the scarcity of large-scale training corpora, these models are validated only on small task-specific datasets, thus scaling and generalizing these models to more languages is non-trivial.

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
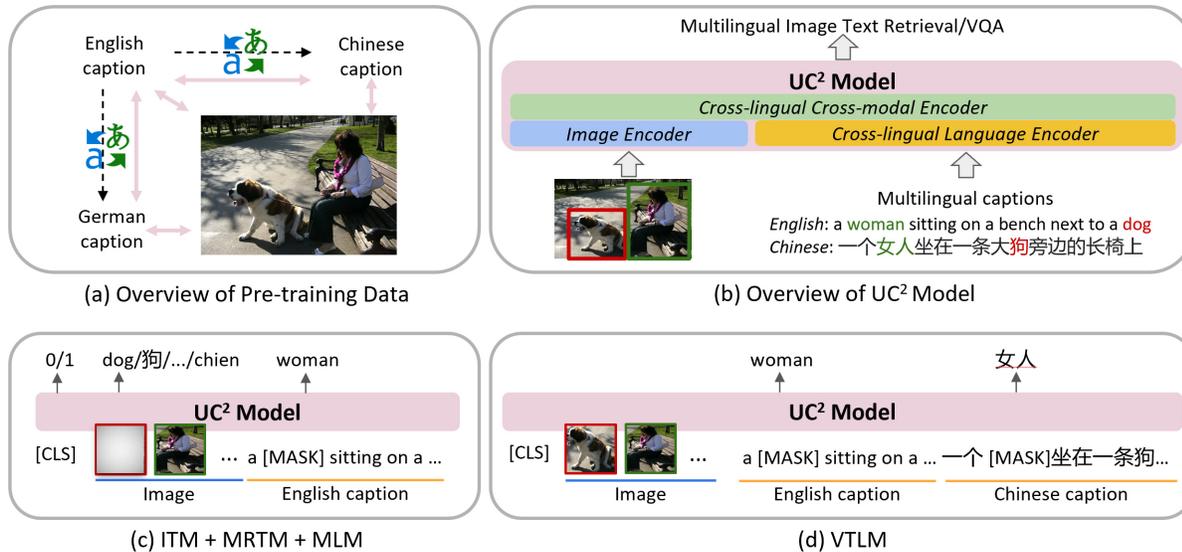
Figure 2. An overview of UC^2 model. Figure (a) shows the construction of multilingual multimodal pre-training corpus via machine translation. (b) depicts the overall UC^2 framework, which is pre-trained with a massive corpus of multilingual caption-image pairs. Figure (c) and (d) illustrate details of four pre-training tasks.

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, UC^2 (**U**niversal **C**ross-lingual **C**ross-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 UC^2

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 UC^2, 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[1] 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 $\boldsymbol{v} = \{v_1, v_2, \cdots, v_m\}$ with Faster R-CNN [39]. For each region, we also extract location features via a 7-dimensional vector: $\boldsymbol{p} = [x_1, y_1, x_2, y_2, w, h, w * h]$, 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^{l_i} in language l_i to BPE tokens $\boldsymbol{t}^{l_i} = \{t_1^{l_i}, t_2^{l_i}, \cdots, t_n^{l_i}\}$ 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:

---

[1]We use Microsoft Azure Translation API Service and will release the translated captions.

Masked Language Modeling (MLM), Image-Text Matching (ITM), Masked Region-to-Token Modeling (MRTM), and Visual Translation Language Modeling (VTLM), as shown in Figure 2 (c) and (d). We continuously optimize our model with the four objectives on multilingual image-text pairs to capture the cross-modal alignment between vision and different languages. As the translated captions are associated to the same image, cross-lingual alignment is also enforced using visual context as the anchor.

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 $v = \{v_1, v_2, \cdots, v_m\}$ and its associated caption words $w^{l_i} = \{w_1^{l_i i}, \cdots, w_T^{l_i}\}$ in language $l_i \in L$, and mask indices as $m \in \mathbb{N}^M$, we randomly mask a word $w_m^{l_i}$ with the probability of 15% and replace the masked word with a special token [mask]. The objective is to predict the masked word $w_m^{l_i}$ based on the surrounding words w_{nm} and all image regions v, by minimizing the negative log-likelihood:

$$\mathcal{L}_{MLM}(\theta) = -\mathbb{E}_{(w^{l_i},v)\sim D} \log P_\theta(w_m^{l_i}|w_{\setminus m}^{l_i}, v),$$

where θ 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 $s_\theta(w^{l_i}, v)$ between 0 and 1, which predicts whether the input image v and the text input w^{l_i} 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:

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

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 (*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 (e.g., "dog", "cat", provided by a pretrained 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 $c_{v_m}^{l_i}$ on the masked region based on the observation of surrounding image regions $v_{\setminus m}$ and the paired caption w^{l_i} in language l_i, by minimizing the negative log-likelihood:

$$\mathcal{L}_{MRTM}(\theta) = -\mathbb{E}_{(w^{l_i},v)\sim D} \log P_\theta(c_{v_m}^{l_i}|w^{l_i}, v_{\setminus m})$$

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 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 $h_{\theta_I}(v_i) \in \mathbb{R}^k$ 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:

$$\mathcal{L}_{EA}(\theta_I) = D_{KL}(g(v_i)||h_{\theta_I}(v_i)),$$

where θ_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 (w^{l_i}, w^{l_j}) 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 $w_m^{l_i}$ from the caption of one language. For the paired caption in the other language l_j, we mask the aligned word tokens $w_k^{l_j}$ 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:

$$\mathcal{L}_{VTLM}(\theta) = -\mathbb{E}_{(w^{l_i}, w^{l_j}, v) \sim D} \log P_\theta(w_m^{l_i}, w_k^{l_j} | w_{\backslash m}^{l_i}, w_{\backslash k}^{l_j}, v)$$

4. Experiments

In this section, we provide detailed experiments to evaluate our proposed UC^2 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 the original MSCOCO dataset by collecting 820K Japanese captions for 165K COCO images. Similarly, Li et al. [34] extend MSCOCO by collecting 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)

Method	Flickr30K				MSCOCO			Meta-Ave
	EN	DE	FR	CS	EN	ZH	JA	
SOTA without pre-training								
EmbN[47]	72.0	60.3	54.8	46.3	76.8	73.2	73.5	65.3
PAR.EmbN [19]	69.0	62.6	60.6	54.1	78.3	76.0	74.8	67.9
S-LIWE [49]	76.3	72.1	63.4	59.4	80.9	73.6	70.0	70.8
MULE [28]	70.3	64.1	62.3	57.7	79.0	75.9	75.6	69.3
SMALR [5]	74.5	69.8	65.9	64.8	81.5	77.5	76.7	73.0
English-only Fine-tune								
M^{3}P[24]	87.4	58.5	46.0	36.8	88.6	53.8	56.0	60.7
UC^2	87.2	74.9	74	67.9	88.1	82	71.7	78.0
Translate-Test								
UNITER$_{CC}$[10]	87.7	81.2	81.9	80.2	88.4	87.3	82.2	84.1
Single-Language Fine-tune								
M^{3}P[24]	87.4	82.1	67.3	65.0	88.6	75.8	80.1	78.0
UC^2	87.2	83.8	77.6	74.2	88.1	84.9	87.3	83.3
All-Language Fine-tune								
M^{3}P[24]	87.7	82.7	73.9	72.2	**88.7**	86.2	**87.9**	82.8
UC^2	**88.2**	**84.5**	**83.9**	**81.2**	88.1	**89.8**	87.5	**86.2**

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.

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.[2] 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.[3]

Implementation Details UC^2 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 warmup 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 batchsize 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 UC^2 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 pretrained UC^2 model.

4.1.1 Evaluation on Multilingual Retrieval

We compare UC^2 with state-of-the-art methods on image retrieval and text retrieval in two different settings:

- **English-only Fine-tune:** We finetune the pre-trained model on just the English training data.
- **Single-Language Fine-tune:** We finetune the pretrained model on training data for each target language.
- **All-Language Fine-tune:** We finetune 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 UNITER$_{CC}$ based on [10], which is pre-trained on Conceptual Conception English data and finetuned on English training data in the downstream tasks. By translating the

---
[2]We only consider top-3129 frequent answers for VQA v2.0 and top-3000 frequent answers for VQA VG Japanese.

[3]BLEU score is used to compute a soft mapping score between the predicted answer and the ground-truth answer, assuming answers with many overlapping words should share similar semantic meaning.

method	VQA v2.0 Test-Dev Acc	VG VQA JA Acc	VG VQA JA BLEU
MCAN [53]	70.63	-	-
PCATT [42]	-	19.2	-
Vil-BERT [36]	70.55	-	-
VL-BERT [44]	71.16	-	-
UNITER$_{CC}$ [10]	71.22	22.7	11.8
UC^2	**71.48**	**34.2**	**26.8**

Table 2. Evaluation results on multilingual VQA task over VQA v2.0 and VG VQA Japanese datasets. We highlight the results for PCATT in blue as they are using different dev/test splits.

test data from other languages to English, UNITER$_{CC}$ 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-pretrained 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 UC^2 model on all-language data, our model still demonstrates a consistent advantage over M^{3}P on the majority of languages, with 3.4% improvement on Meta-Ave. Our best model is also better than the strong translate-test baseline UNITER$_{CC}$ 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 splitted 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 UNITER$_{CC}$ 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 UC^2. 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 (UNITER$_{CC}$) performs much worse than UC^2 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 UC^2 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 finetuned 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 improve-

Pretraiing Objective	Flickr30K				MSCOCO				VQA v2.0	VG VQA JA	
	EN	DE	FR	CS	EN	ZH	JA	Meta-Ave	Test-Dev Acc	Acc	BLEU
UC^2 (full model)	**88.2**	**84.5**	**83.9**	**81.2**	**88.1**	**89.8**	**87.5**	**86.2**	**71.48**	**34.2**	**26.8**
-VTLM	87.5	83.6	82.4	79.6	87.7	89.2	87.2	85.3	71.45	34.1	26.7
-MRTM	87.6	83.7	82.0	80.0	87.9	89.4	87.4	85.4	70.93	33.5	26.4
-VTLM-MRTM	86.8	82.9	81.3	79.3	87.5	88.9	86.7	84.8	69.94	33.4	26.4

Table 3. Ablation study on pre-training objectives.

Topology	Flickr30K				MSCOCO			
	EN	DE	FR	CS	EN	ZH	JA	Meta-Ave
UC^2 (Image pivoting)	**87.5**	**83.6**	**82.4**	**79.6**	87.7	**89.2**	87.2	**85.3**
UC^2 (English pivoting)	86.2	81.9	80.7	77.4	**88.1**	88.5	**87.3**	84.2

Table 4. Comparison between the pre-training topology of pivoting on image against pivoting on English.

ment 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 (similar to M3P). In this setting, we train UC2 with all the pre-training objectives on English Conceptual Caption data as in our full model, except for VTLM which involves image as one of the pivoting points. To capture the alignment between English and other languages, we train UC^2 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 UC^2 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 UC^2, 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

Figure 3. Visualization of Text-to-Image Attention on aligned words across English, German and Czech (Flickr30K).

words, it also connects different languages by grounding them to similar image regions.

5. Conclusion

We present the first MT-augmented pre-training model UC^2 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] Roee Aharoni, Melvin Johnson, and Orhan Firat. Massively multilingual neural machine translation. In *North American Chapter of the Association for Computational Linguistics*, 2019. 1

[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

[5] Andrea Burns, Donghyun Kim, Derry 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

[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. 1

[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

[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

[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. 5, 6, 7

[21] Jindřich Helcl 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

translations via images with a massively multilingual image dataset. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2566–2576, 2018. 1

[23] Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalization. In *International Conference on Machine Learning*, 2020. 1, 3

[24] Haoyang Huang, Lin Su, Di Qi, Nan Duan, Edward Cui, Taroon Bharti, Lei Zhang, Lijuan Wang, Jianfeng Gao, Bei Liu, Jianlong Fu, Dongdong Zhang, Xin Liu, and Ming Zhou. M3p: Learning universal representations via multi-task multilingual multimodal pre-training, 2020. 1, 5, 6

[25] Po-Yao Huang, Junjie Hu, Xiaojun Chang, and Alexander Hauptmann. Unsupervised multimodal neural machine translation with pseudo visual pivoting. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8226–8237, Online, July 2020. Association for Computational Linguistics. 3

[26] Zhicheng Huang, Zhaoyang Zeng, Bei Liu, Dongmei Fu, and Jianlong Fu. Pixel-bert: Aligning image pixels with text by deep multi-modal transformers, 2020. 1, 2

[27] Andrej Karpathy and Fei-Fei Li. Deep visual-semantic alignments for generating image descriptions. *CoRR*, abs/1412.2306, 2014. 5

[28] Donghyun Kim, Kuniaki Saito, Kate Saenko, Stan Sclaroff, and Bryan A. Plummer. MULE: Multimodal Universal Language Embedding. In *AAAI Conference on Artificial Intelligence*, 2020. 1, 3, 6

[29] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. 6

[30] 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. Visual genome: Connecting language and vision using crowdsourced dense image annotations. 2016. 6

[31] Taku Kudo and John Richardson. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium, Nov. 2018. Association for Computational Linguistics. 3

[32] Guillaume Lample and Alexis Conneau. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems*, 2019. 1, 8

[33] Gen Li, Nan Duan, Yuejian Fang, Ming Gong, Daxin Jiang, and Ming Zhou. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training, 2019. 2, 4

[34] Xirong Li, Xiaoxu Wang, Chaoxi Xu, Weiyu Lan, Qijie Wei, Gang Yang, and Jieping Xu. COCO-CN for cross-lingual image tagging, captioning and retrieval. *CoRR*, abs/1805.08661, 2018. 2, 5

[35] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. Oscar: Object-semantics aligned pre-training for vision-language tasks, 2020. 1, 2, 5

[36] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks, 2019. 1, 2, 4, 7

[37] Jiasen Lu, Vedanuj Goswami, Marcus Rohrbach, Devi Parikh, and Stefan Lee. 12-in-1: Multi-task vision and language representation learning, 2020. 2

[38] Janarthanan Rajendran, Mitesh M. Khapra, Sarath Chandar, and Balaraman Ravindran. Bridge correlational neural networks for multilingual multimodal representation learning. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 171–181, San Diego, California, June 2016. Association for Computational Linguistics. 3

[39] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. Faster R-CNN: towards real-time object detection with region proposal networks. *CoRR*, abs/1506.01497, 2015. 3, 4

[40] Ramon Sanabria, Ozan Caglayan, Shruti Palaskar, Desmond Elliott, Loïc Barrault, Lucia Specia, and Florian Metze. How2: A large-scale dataset for multimodal language understanding. *CoRR*, abs/1811.00347, 2018. 8

[41] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, Melbourne, Australia, July 2018. Association for Computational Linguistics. 1, 3, 5

[42] Nobuyuki Shimizu, Na Rong, and Takashi Miyazaki. Visual question answering dataset for bilingual image understanding: A study of cross-lingual transfer using attention maps. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1918–1928, Santa Fe, New Mexico, USA, Aug. 2018. Association for Computational Linguistics. 6, 7

[43] Gunnar A. Sigurdsson, Jean-Baptiste Alayrac, Aida Nematzadeh, Lucas Smaira, Mateusz Malinowski, Joao Carreira, Phil Blunsom, and Andrew Zisserman. Visual grounding in video for unsupervised word translation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 3

[44] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. Vl-bert: Pre-training of generic visual-linguistic representations, 2020. 2, 4, 7

[45] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers, 2019. 1, 2

[46] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Infor-*

mation Processing Systems, volume 30, pages 5998–6008. Curran Associates, Inc., 2017. 3

[47] Liwei Wang, Yin Li, and Svetlana Lazebnik. Learning two-branch neural networks for image-text matching tasks. *CoRR*, abs/1704.03470, 2017. 6

[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. 8

[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. 1, 3, 6

[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. 1, 3

[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. 2, 5

[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. 5

[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. 7

[54] Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason J. Corso, and Jianfeng Gao. Unified vision-language pretraining for image captioning and vqa. In *AAAI*, 2020. 1, 2

[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. 1, 3