

Cross-View Language Modeling: Towards Unified Cross-Lingual Cross-Modal Pre-training

Yan Zeng^{1*†} Wangchunshu Zhou^{1*} Ao Luo² Xinsong Zhang¹

¹ByteDance AI Lab ²Beijing Normal University

Abstract

In this paper, we introduce Cross-View Language Modeling, a simple and effective language model pre-training framework that unifies cross-lingual cross-modal pre-training with shared architectures and objectives. Our approach is motivated by a key observation that cross-lingual and cross-modal pre-training share the same goal of aligning two different views of the same object into a common semantic space. To this end, the cross-view language modeling framework considers both multi-modal data (i.e., image-caption pairs) and multi-lingual data (i.e., parallel sentence pairs) as two different views of the same object, and trains the model to align the two views by maximizing the mutual information between them with conditional masked language modeling and contrastive learning. We pre-train **CCLM**³, a Cross-lingual Cross-modal Language Model, with the cross-view language modeling framework. Empirical results on IGLUE, a multi-lingual multi-modal benchmark, and two multi-lingual image-text retrieval datasets show that while conceptually simpler, CCLM significantly outperforms the prior state-of-the-art with an average absolute improvement of over 10%. Notably, CCLM is the first multi-lingual multi-modal model that surpasses the translate-test performance of representative English vision-language models by zero-shot cross-lingual transfer.

1 Introduction

Recently, the tremendous success of self-supervised language model pre-training [1–15] has been expanded to the multi-lingual [16–19] and multi-modal [20–24] domain. Advances on multi-lingual pre-training enables cutting-edge language technology to benefit a much boarder group of users including non-English speakers. Similarly, multi-modal pre-training makes pre-trained models applicable to a much larger set of tasks and user groups. Both of these directions make people’s lives in a multi-lingual multi-modal world easier. Therefore, a natural next step is to explore multi-lingual multi-modal pre-training which enables pre-trained models to solve multi-modal tasks expressed in non-English languages without the need of collecting training data in these languages, which can be very costly for certain low-resource languages.

While appealing, multi-lingual multi-modal pre-training has its own challenges. Unlike multi-lingual pre-training and multi-modal pre-training where relatively large amount of parallel data is available, there exists only a few multi-lingual multi-modal corpora and their language coverage is also limited. Two pioneering works, M³P [25] and UC² [26], propose to pivot either on English texts or images to align multi-lingual multi-modal representations. Both of them introduce a number of new objectives to make use of the anchor for alignment. However, a recent benchmark on multi-lingual multi-modal pre-training [27] reveals that these multi-lingual multi-modal pre-trained models are still falling short: while achieving seemingly promising zero-shot cross-lingual transfer performance on some

Preprint. Under review.

*Equal Contribution.

†Correspondence to: <zengyan.yanne@bytedance.com>.

³The code and pre-trained models are available at <https://github.com/zengyan-97/CCLM>.

vision-and-language tasks, they still significantly under-perform “translate-test”, a simple baseline which translates the test examples into English and uses an English-only vision-language model for inference. This prevents existing multi-lingual multi-modal models to be applicable in real-world applications. In contrast, multi-lingual pre-trained models such as XLM-R [17] significantly outperforms the translate-test baseline in most languages and is widely used in practical applications.

This paper aims to fully exploit the potential of multi-lingual multi-modal pre-training. We point out two major limitation of current state-of-the-arts. First, existing methods do not exploit parallel text corpora, which can be easily collected and are abundant for many language pairs. Instead, M³P performs masked language modeling with monolingual texts in different languages for multi-lingual alignment. However, parallel texts are shown to be more helpful according to multi-lingual pre-training literature [17, 19]. Second, a number of new pre-training objectives involving specific architecture changes and different input-output formats are introduced for English or image pivoting, making it non-trivial to combine them together for better performance and scale to larger data.

In this work, we argue that multi-lingual and multi-modal pre-training are essentially achieving the same goal of aligning two different views of a same object into a common semantic space. Therefore, we believe these two seemingly different strategies can be combined into a unified framework. To this end, we introduce cross-view language modeling, a simple and effective framework that unifies cross-lingual and cross-modal pre-training with shared architecture and objectives. Specifically, we consider both multi-modal data (i.e., image-caption pairs) and multi-lingual data (i.e., parallel sentence pairs) as pairs of two different views of the same object. With either multi-modal or multi-lingual data as input, we encode the two views with Transformer models and then fuse their representations with a cross-attention Transformer model shared for both cross-modal and cross-lingual fusion. We train the model to align the two views into a common semantic space by maximizing the mutual information between them with a conditional masked language modeling objective, a contrastive learning objective, and a matching objective. In this way, the cross-view language modeling framework unifies English pivoting and image pivoting schemes seamlessly and makes the best of both worlds.

To evaluate the effectiveness of our approach, we pre-train CCLM, a Cross-lingual Cross-modal Language Model, with the proposed cross-view language modeling framework. Experimental results show that CCLM significantly outperforms prior state-of-the-art with an averaged absolute improvement of 11.4% and 32.7% on multi-lingual vision-language understanding and retrieval tasks in terms of accuracy and R@1 on IGLUE [27], a recently released multi-lingual multi-modal benchmark. Notably, CCLM is the first multi-lingual vision-language model that surpasses the “translate-test” performance of mono-lingual vision-language models via zero-shot cross-lingual transfer, which we believe is a crucial step towards practical multi-lingual multi-modal pre-training.

Contributions. (1) We propose a cross-view language modeling framework that unifies multi-lingual and multi-modal pre-training with shared architectures and objectives. (2) We pre-train CCLM with the proposed approach on public available image-text pairs and parallel sentence pairs. (3) CCLM advances the state-of-the-art of multi-lingual vision-language pre-training by a large margin and surpass the translate-test baseline for the first time.

2 Related Work

Multi-lingual Pre-training Multilingual BERT [3] demonstrates that good cross-lingual transfer results can be achieved by performing masked language modeling on multi-lingual corpora with shared vocabulary and weight. Later, XLM [16], XLM-R [17], and Unicoder [28] introduce a number of new objectives including translation language modeling (TLM), cross-lingual word recovery, and cross-lingual paraphrase classification to improve multi-lingual pre-training. More recently, MAD-X [18] and InfoXLM [19] further improve multi-lingual pre-training via adapter [29] and contrastive learning.

Vision-Language Pre-training Inspired by the success of language model pre-training, a number of work [20, 21, 24, 23, 30] investigates vision-language pre-training on large scale image-caption pairs and proposes a number of objectives to align vision and language representations, including masked multi-modal modeling, multi-modal alignment prediction, RoI feature regression, image-text matching, to name a few. Vision-language pre-training has reshaped the landscape of vision-and-language research and pushed the state-of-the-arts on a wide range of vision-language tasks [31].

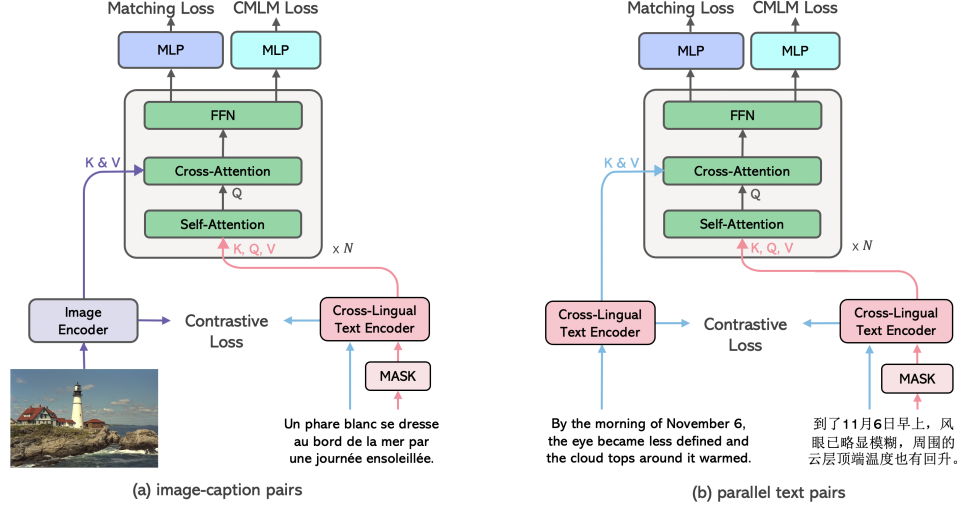


Figure 1: Illustration of the cross-view language modeling framework. CCLM takes two different views of the same object, i.e., either (a) image-caption pairs or (b) parallel sentence pairs, as input. CCLM first encodes the two views separately with Transformer encoders. Then the representations of the two views are fused by a Transformer-based fusion model, which is shared for both cross-lingual and cross-modal fusion. CCLM is optimized by maximizing the mutual information between the two views via conditional masked language modeling loss, contrastive loss, and matching loss.

However, it is non-trivial to collect large scale image-caption pairs in other languages. As such, most existing vision-language pre-trained models are limited to English tasks.

Multi-lingual Multi-modal Pre-training Multi-lingual multi-modal pre-training aims to make multi-modal models applicable on non-English texts by cross-lingual transfer. In this paper we mainly consider multi-modal in the vision-language context. The key difficulty of multi-lingual multi-modal pre-training is the lack of non-English image-text pairs. Two representative works tackle the lack of non-English image-text pairs by pivoting on either English texts or images. Specifically, M³P [25] uses English as pivot and alternates between English-only vision-language pre-training, multi-lingual masked language modeling, and multimodal code-switched training. UC² [26], on the other hand, translate English captions into multiple languages and considers images as the anchor, achieving state-of-the-art on various multi-lingual vision-language tasks. More recently, MURAL [32] pre-trains a dual encoder model on massive multi-lingual multi-modal data via contrastive learning and achieves new state-of-the-art on multi-lingual image-text retrieval datasets. However, the dual encoder architecture of MURAL makes it unable to perform multi-modal understanding tasks well.

3 Cross-View Language Modeling

3.1 Overview

Cross-view language modeling is a simple framework that unifies cross-lingual pre-training and cross-modal pre-training with shared architecture and objectives. CCLM consists of an image encoder, a cross-lingual text encoder, and a fusion model. All components are Transformer-based. Specifically, the image encoder [33] first splits an image into non-overlapping patches, and then embeds these patches with transformer layers, yielding $\{\vec{v}_{cls}, \vec{v}_1, \dots, \vec{v}_{N_1}\}$. For an image of resolution of 224x224 and patch size of 32x32, we have $N_1 = 49$. Similarly, the cross-lingual text encoder encodes a text input via transformer layers, yielding $\{\vec{w}_{cls}, \vec{w}_1, \dots, \vec{w}_{N_2}\}$. N_2 is the length of the text input. Then, the fusion model fuses text features with the corresponding image features or features of the translated text based on cross-attention, producing $\{\vec{x}_{cls}, \vec{x}_1, \dots, \vec{x}_{N_2}\}$.

As illustrated in Figure 1, with either (text, image) pairs or (text, translation) pairs as input, we consider the paired input as two different views and train the model to align their representations in a common semantic space. This unified cross-view perspective allows us to share input-output

formats, architectures, and training objectives between cross-lingual inputs and cross-modal inputs. Specifically, we completely share the fusion model for both cross-lingual fusion and cross-modal fusion, and optimize the model by contrastive loss, matching loss, and conditional masked language modeling loss for both cross-lingual and cross-modal inputs. We select these objectives because they are universally effective in both cross-lingual and cross-modal pre-training literature [19, 34]. We will show that the three loss maximize sequence-level and token-level mutual information between image-caption pairs or parallel sentence pairs. On the other hand, we empirically find that the three loss are more effective for cross-lingual cross-modal pre-training than certain task-specific loss such as masked region-to-token language modeling which is specially for multi-modal pre-training or translation language modeling for multilingual pre-training.

3.2 A Mutual Information Maximization Perspective

In this section, we explain our approach from an information-theoretic perspective. Formally, given two random variables A and B , mutual information $I(A, B)$ measures dependencies between the two random variables. We define $A = a$ and $B = b$ as two different views of a data point, which can be either an image-caption pair or a parallel sentence pair. In this case, we will show that CCLM maximizes a lower bound of $I(A, B)$ for cross-lingual cross-modal pre-training by minimizing the InfoNCE loss [35] defined as:

$$\mathcal{L}_{\text{nce}} = -\mathbb{E}_{p(A, B)} \left[\log \frac{\exp(f_{\theta}(a, b))}{\sum_{\tilde{b} \in \tilde{B}} \exp(f_{\theta}(a, \tilde{b}))} \right], \quad (1)$$

where $f_{\theta} \in \mathbb{R}$ is a function parameterized by θ and \tilde{B} contains the positive sample b and $|\tilde{B}| - 1$ negative samples.

The contrastive loss between the image encoder and the cross-lingual text encoder is a symmetric version of \mathcal{L}_{nce} :

$$\mathcal{L}_{\text{cl}} = -\frac{1}{2} \mathbb{E}_{p(A, B)} \left[\log \frac{\exp(f_{\theta}(a, b))}{\sum_{\tilde{b} \in \tilde{B}} \exp(f_{\theta}(a, \tilde{b}))} + \log \frac{\exp(f_{\theta}(a, b))}{\sum_{\tilde{a} \in \tilde{A}} \exp(f_{\theta}(\tilde{a}, b))} \right], \quad (2)$$

where $|\tilde{A}| = |\tilde{B}| = N$ is the batch size, and we predict (a, b) pairs from in-batch negatives. $f_{\theta}(a, b) = g_v(\vec{v}_{\text{cls}})^{\top} g_w(\vec{w}_{\text{cls}}) / \tau$ given an image-caption pair or $f_{\theta}(a, b) = g_w(\vec{w}_{\text{cls}}^a)^{\top} g_v(\vec{v}_{\text{cls}}^b) / \tau$ given a translation pair. \vec{v}_{cls} and \vec{w}_{cls} are the output [CLS] embedding of the image encoder⁴ and the cross-lingual text encoder. g_v and g_w are transformations that map the [CLS] embeddings to normalized lower-dimensional representations. τ is a learnable temperature parameter.

Similarly, the matching loss applied on the output [CLS] embedding of the fusion model (denoted as $\vec{x}_{\text{cls}}(a, b)$) can also be viewed as a symmetric version of \mathcal{L}_{nce} :

$$\mathcal{L}_{\text{match}} = -\frac{1}{2} \mathbb{E}_{p(A, B)} \left[\log \frac{\exp(f_{\theta}(a, b))}{\exp(f_{\theta}(a, b)) + \exp(f_{\theta}(a, b_{\text{neg}}))} + \log \frac{\exp(f_{\theta}(a, b))}{\exp(f_{\theta}(a, b)) + \exp(f_{\theta}(a_{\text{neg}}, b))} \right], \quad (3)$$

where we only sample a negative instance for each ground-truth (a, b) pair and predict whether a pair is matched (true or false). In this case, $f_{\theta}(a, b) = \vec{v}_{\text{true}}^{\top} \vec{x}_{\text{cls}}(a, b)$, where \vec{v}_{true} is a parametric vector.

The conditional MLM loss can also be interpreted as maximizing mutual information [36] between the context $c = (\hat{a}, b)$ (\hat{a} denotes the masked text input, and b is the corresponding image or translated text) and the masked token w_i in a :

$$\mathcal{L}_{\text{mlm}} = -\mathbb{E}_{p(C, W)} \left[\log \frac{\exp(f_{\theta}(c, w_i))}{\sum_{\tilde{w} \in \mathcal{V}} \exp(f_{\theta}(c, \tilde{w}))} \right], \quad (4)$$

where $f_{\theta}(c, w_i) = \psi(w_i)^{\top} \vec{x}_i(\hat{a}, b)$. \vec{x}_i is the output vector at w_i position of the fusion model. $\psi(w) : \mathcal{V} \rightarrow \mathbb{R}^d$ is a lookup function that maps a word token w into a parametric vector. \mathcal{V} is the full vocabulary set.

Finally, the pre-training objective of CCLM is defined as:

$$\mathcal{L} = \mathcal{L}_{\text{cl}} + \mathcal{L}_{\text{match}} + \mathcal{L}_{\text{mlm}}, \quad (5)$$

⁴Some vision transformers, e.g. Swin-Transformer, use the output vector of average pooling layer as the [CLS] embedding.

where the contrastive loss and matching loss maximize sequence-level mutual information while the MLM loss maximizes token-level mutual information, which are complement of each other.

4 Experiment

4.1 Experimental Settings

4.1.1 Pre-training Datasets

Multi-modal Data We pre-train CCLM on the combination of image-caption pairs and parallel texts. For image-caption pairs, we follow the practice of UC² and use their released translation-augmented version of CC3M dataset. It contains the original CC3M image-caption pairs [37] and machine-translated captions in five different languages (German, French, Czech, Japanese, and Chinese). In addition to this setting, we also experiment with another setting by including the COCO dataset [38] and the Visual Genome (VG) dataset [39], which contains around 1 million image-caption pairs in total. We consider this setup because COCO and VG datasets are commonly used in vision-language pre-training literature while not used by previous work on multi-lingual multi-modal pre-training. We denote our model trained with these two variants as CCLM_{3M} and CCLM_{4M} respectively.

Multi-lingual Data As for parallel text corpus, we collect a subset of the WikiMatrix [40] dataset containing parallel texts between English and other languages in the IGLUE benchmark⁵. The multi-lingual pre-training data consists of 19M parallel sentence pairs in total.

4.1.2 Implementation Details

We initialize the image encoder by Swin Transformer [41] which consists of 12 Transformer layers. The cross-lingual text encoder and the fusion model are initialized with the first and second half of XLM-R [42] respectively, consisting of six layers each. The image encoder takes images of resolution of 224×224 as input. The maximum sequence length is set to 30 and 64 for image-caption pairs and parallel texts respectively. During fine-tuning, we increase the image resolution to 384×384 and interpolate the positional embeddings of image patches following Dosovitskiy et al. [33].

We apply mixed precision for pre-training. Following UC², we train the model for 30 epochs on 8 NVIDIA A100 GPUs and the batch size is set to 1024, which takes ~ 1.5 days. We use the AdamW [43] optimizer with a weight decay of 0.02. The learning rate is warmed-up to $1e^{-4}$ from $1e^{-5}$ in the first 2500 steps and decayed to $1e^{-5}$ following a linear schedule. The pre-training is alternating between image-caption batches and parallel text batches.

4.1.3 Downstream Tasks

We adapt CCLM to two series of downstream datasets including IGLUE benchmark, a recently released benchmark for evaluating multi-lingual multi-modal pre-training, and multi-lingual image-text retrieval datasets including the multi-lingual version of Flickr30K [44, 45] and MSCOCO [46]. We describe the details of downstream datasets as follows.

Flickr30K: This dataset extended Flickr30K [44] from English (en) to German (de), French (fr) and Czech (cs). It contains 31,783 images and provides five captions per image in English and German, and one caption per image in French and Czech. Dataset splits are defined as the original Flickr30K.

MSCOCO: This dataset extends the MSCOCO caption dataset [46] by translating the captions into Japanese [47] and Chinese [48]. The Japanese and Chinese subsets consist of 820k and 20k captions respectively. Following previous work, we use the same train, dev, and test splits for English and Japanese as defined in Karpathy and Li [49]. As for Chinese, we use the COCO-CN split [48].

XVNL: The Cross-lingual Visual NLI dataset is released by the IGLUE benchmark. It is collected by combining SNLI [50] with its multi-modal [51] and multi-lingual [52] counterparts. It requires the model to predict if a text-hypothesis “entails”, “contradicts”, or is “neutral” to an image-premise.

xGQA: The Cross-lingual Grounded Question Answering task [53] is collected by manually translating the GQA [54] validation set into 7 languages. It requires a model to answer several types of structured questions about an image. We model GQA as a generation task following Li et al. [34].

⁵Exact languages and numbers of parallel sentences per language are listed in Table 4 in the Appendix

MaRVL: The Multicultural Reasoning over Vision and Language dataset [55] requires to determine whether a textual description is true or false about a pair of images. The MaRVL dataset is used for testing and the NLVR2 [56] dataset is used for training.

xFlickr&CO and WIT: The xFlickr&CO dataset is collected by combining 1000 images from Flickr30K and MSCOCO respectively and crowdsource image descriptions in 6 other languages. Similarly, the Wikipedia-based Image Text dataset [57] is collected from Wikipedia in 108 languages. We follow the data preprocessing and splitting details in IGLUE for both datasets.

For all retrieval tasks, we apply the same ranking strategy as ALBEF [34] for inference. For all experiments on Flickr30K and MSCOCO, we fine-tune the model for 10 epochs with a batch size of 160 on 8 GPUs, a learning rate of $1e^{-5}$, a weight decay of 0.01, and a warmup ratio of 0.1. We find these hyperparameters work well across settings, datasets, and languages. Therefore, we do not perform any hyperparameter search on Flickr30K and MSCOCO. For IGLUE datasets, we present detailed hyperparameters for fine-tuning in Table 5 and Table 6 in the Appendix.

4.1.4 Compared Models

mUNITER and xUNITER: A multi-lingual variant of the UNITER [23] model pre-trained by Liu et al. [55]. The model is pre-trained by alternating between a batch of multi-modal English data from CC3M with UNITER objectives and a batch of text-only multilingual Wikipedia data with the MLM objective. mUNITER and xUNITER differ in their initialization: mUNITER and xUNITER are initialized from mBERT and XLM-R.

M³P: A multi-lingual multi-modal model initialized from XLM-R and pre-trained with the combination of multilingual masked language modeling, multi-modal code-switched masked language modeling, multi-modal code-switched masked region modeling, and multi-modal code-switched visual-linguistic matching. The code-switched training method allows the model to explicitly align images with non-English languages. In each multi-modal batch, image-text pairs are fed to the model either fully in English or with code-switched words according to a given sampling ratio. Similar to mUNITER and xUNITER, the model is trained by alternating multi-modal and multi-lingual batches.

UC²: The state-of-the-art multi-lingual vision-language model which relies on (text-only) machine translation technologies to obtain CC3M data in five languages (Czech, French, German, Japanese, and Mandarin). The model is then pre-trained on multi-lingual multi-modal batches where a caption is sampled uniformly from the available languages for each image. As for pre-training objectives. In addition to conventional vision-language pre-training objectives, a visual-conditioned translation language modeling objective is added to improve multi-lingual multi-modal alignment.

All compared models are pre-trained with multi-modal data from CC3M. mUNITER, xUNITER, and M³P use Wikipedia data in different languages as multi-lingual data while UC² uses translated version of CC3M as multi-lingual data.

4.2 Experimental Results

4.2.1 Results on IGLUE Benchmark

We first evaluate CCLM on the IGLUE benchmark. We follow the practice of IGLUE and report results in both zero-shot and few-shot cross-lingual transfer settings. In the zero-shot setting, the models fine-tuned on English train sets are directly evaluated on target languages. In the few-shot setting, the English trained models are continually fine-tuned with a few labeled examples in a target language before evaluating on this language. We select exactly the same few-shot examples following IGLUE instructions to ensure our results are compatible with that reported in IGLUE. The results are shown in Table 1. Results of compared models are copied from IGLUE. We omit few-shot evaluation on the WIT dataset because this setup is also omitted in IGLUE.

First, for zero-shot cross-lingual transfer results, we can see that CCLM_{3M} outperforms all compared models by a substantial margin while pre-trained on the same multi-modal data. Specifically, compared to UC², the prior state-of-the-art, CCLM_{3M} obtains an average accuracy improvement of 11.4% on multi-lingual multi-modal understanding tasks including XVNLI, xGQA, and MaRVL, and an average R@1 improvement of 47.3% and 18.2% on multi-lingual multi-modal retrieval datasets including xFlickr&CO and WIT. This confirms that previous multi-lingual multi-modal models fail to fully exploit the potential of multi-lingual multi-modal pre-training and our proposed cross-view

Model	NLI	QA	Reasoning	Retrieval			
	XVNLI	xGQA	MaRVL	xFlickr&CO		WIT	
				IR	TR	IR	TR
<i>Fine-tune model on English training set (Zero-Shot)</i>							
mUNITER	53.69	9.97	53.72	8.06	8.86	9.16	10.48
xUNITER	58.48	21.72	54.59	14.04	13.51	8.72	9.81
M ³ P	58.25	28.17	56.00	12.91	11.90	8.12	9.98
UC ²	62.05	29.35	57.28	20.31	17.89	7.83	9.09
CCLM _{3M}	74.64(.59)	42.36(.68)	65.91(.40)	67.35(.31)	65.37(.10)	27.46(.14)	28.66(.19)
CCLM _{4M}	73.32(.24)	46.24(.21)	67.17(.42)	76.56(.14)	73.46(.09)	27.48(.18)	28.75(.23)
<i>Few-shot train English fine-tuned model on target languages (Few-Shot)</i>							
mUNITER	53.95	37.21	53.41	8.54	9.32	-	-
xUNITER	60.55	40.68	57.46	14.30	13.54	-	-
M ³ P	59.36	41.04	49.79	13.21	12.26	-	-
UC ²	63.68	42.95	58.32	19.79	17.59	-	-
CCLM _{3M}	75.15(.03)	50.94(.02)	70.53(.18)	66.04(.05)	68.15(.04)	-	-
CCLM _{4M}	74.19(.07)	55.11(.00)	72.57(.48)	74.74(.01)	75.85(.03)	-	-
<i>Translate everything to English and use English-only model (Translate-Test)</i>							
UNITER	73.65	50.62	61.92	41.04	37.49	15.43	16.01
ViLBERT	73.45	50.33	62.39	36.97	33.21	15.40	16.93
VisualBERT	74.12	48.72	62.35	41.64	36.44	15.36	15.75
VL-BERT	73.86	49.78	64.16	38.18	31.84	15.11	16.09

Table 1: **Results on IGLUE benchmark.** R@1 and Accuracy are reported for retrieval tasks (xFlickr&CO and WIT) and understanding tasks (XVNLI, xGQA, MaRVL) respectively. For our models, mean and standard deviation (in brackets) of 3 different runs with different random seeds are reported. Results of compared models are directly copied from the IGLUE benchmark.

language modeling framework can better align multi-lingual multi-modal representations with unified objectives. We also find that the performance can be further improved (CCLM_{4M}) by adding COCO and VG data, which are used for pre-training most English VLMs. Notably, CCLM is the first multi-lingual multi-modal pre-trained model that performs competitively with the translate-test results of representative English VLMs tested in the IGLUE benchmark. Concretely, CCLM_{4M} outperforms the translate-test results of all representative English VLMs in the IGLUE benchmark on XVNLI, MaRVL, xFlickr&CO, and WIT while performs slightly worse on the xGQA dataset. This, for the first time, proves the potential of multi-lingual multi-modal pre-training on building practical real-world applications involving vision-language tasks in different languages.

As for few-shot results, we find that similar to existing models, CCLM can also benefit from few-shot learning with a few examples in the target languages. By training on a few examples per class in the target languages, CCLM consistently outperforms translate-test results of English VLMs with a larger margin. This further confirms the potential of multi-lingual multi-modal pre-training.

4.2.2 Results on Multi-lingual Retrieval

We also compare CCLM with state-of-the-art methods on conventional image retrieval and text retrieval tasks on which UC² and M³P originally reported their results. We follow the practice of prior work and evaluate in three different settings including English-only fine-tuning, single-language fine-tuning, and all-language fine-tuning, where the model is fine-tuned on English data, target language data, and the combination of training data in all languages, respectively.

The results are shown in Table 2. For zero-shot cross-lingual transfer, we can see that CCLM_{3M} also substantially outperforms UC², the prior state-of-the-art, with an averaged improvement of over 16% (in terms of averaged recall) across five languages. This confirms that our approach can better align multi-lingual multi-modal representations. Including COCO and VG data also yields some improvements, which is consistent with previous results on the IGLUE benchmark. Fine-tuning on target languages or the combination of all languages yields consistent improvements. The improvements are not as large as that for UC² and M³P, which is probably because the zero-shot cross-lingual transfer ability of CCLM is strong enough and the performance of our models is already

Model	Flickr30K				MSCOCO		
	EN	DE	FR	CS	EN	ZH	JA
<i>English-only Fine-tune</i>							
M ³ P	87.4	58.5	46.0	36.8	88.6	53.8	56.0
UC ²	87.2	74.9	74.0	67.9	88.1	82.0	71.7
CCLM _{3M}	94.8(.11)	90.3(.08)	90.9(.38)	89.4(.21)	93.2(.05)	91.0(.18)	88.8(.06)
CCLM _{4M}	95.7(.03)	91.7(.12)	92.9(.12)	91.7(.40)	94.3(.07)	91.7(.16)	91.5(.08)
<i>Single-Language Fine-tune</i>							
M ³ P	87.4	82.1	67.3	65.0	88.6	75.8	80.1
UC ²	87.2	83.8	77.6	74.2	88.1	84.9	87.3
CCLM _{3M}	94.8(.11)	91.9(.16)	90.6(.18)	88.9(.05)	93.2(.05)	90.2(.24)	93.3(.26)
CCLM _{4M}	95.7(.03)	93.0(.15)	92.3(.17)	91.5(.06)	94.5(.07)	92.6(.14)	94.2(.10)
<i>All-Language Fine-tune</i>							
M ³ P	87.7	82.7	73.9	72.2	88.7	86.2	87.9
UC ²	88.2	84.5	83.9	81.2	88.1	89.8	87.5
MURAL _{base}	92.2	88.6	87.6	84.2	88.6	-	88.4
MURAL _{large}	93.8	90.4	89.9	87.1	92.3	-	91.6
CCLM _{3M}	95.3(.20)	92.4(.12)	92.1(.15)	91.2(.18)	93.1(.08)	92.2(.16)	93.2(.17)
CCLM _{4M}	96.0(.19)	93.3(.14)	93.7(.07)	92.8(.16)	94.1(.14)	93.0(.07)	94.3(.01)

Table 2: **Results on multi-lingual image-text retrieval.** We compute the average Recall@K for both image-to-text retrieval and text-to-image retrieval with K = 1, 5, 10, as the evaluation metric. For our models, mean and standard deviation (in brackets) of 3 different runs with different random seeds are reported. Results of compared models are directly copied from the corresponding papers. Numbers of MURAL are in gray because it is a dual encoder model and is pre-trained on much larger data, thus not comparable with other results.

saturating. Nevertheless, CCLM_{3M} still substantially outperforms prior state-of-the-art by 9.4% and 6.8% averaged recall across five languages when fine-tuned on target languages or the combination of all languages, respectively. Moreover, CCLM_{4M} also significantly outperforms MURAL_{large}, the prior state-of-the-art in the all-language fine-tuning setting by 3.8% averaged recall across four languages. This is notable because MURAL_{large} is larger than our model and is pre-trained on much more data ($\sim 450\times$ more image-text pairs and $390\times$ more parallel sentence pairs). Moreover, we show that CCLM also outperforms MURAL_{large} in the zero-shot setting (w/o fine-tune) in Table 7.

4.2.3 Cross-lingual Transfer Gap

In addition to absolute cross-lingual transfer results, we also compare the cross-lingual transfer gap of different models. We visualize the ratio of a model’s performance on non-English languages to its performance on English test set, in Figure 2. A larger radar chart indicates the model has a smaller relative transfer gap and can better transfer its performance to non-English test sets. We can see that CCLM’s relative cross-lingual transfer gap is consistently smaller than that of UC² across all tasks in the IGLUE benchmark (a) and all languages in the multi-lingual retrieval datasets (b). The absolute cross-lingual

transfer gap is even more significant. For example, in Table 2, we can see that for M³P, the absolute zero-shot cross-lingual transfer gap between EN-CS and EN-JA in Flickr30K and MSCOCO are 41.4% and 32.6% respectively. This indicates that masked language modeling on unpaired texts in multiple languages are not very effective for cross-lingual alignment of multi-modal models. The gap for UC² is reduced to 13.2% and 16.4%, demonstrating the effectiveness of using machine-translated captions for multi-lingual multi-modal pre-training. Surprisingly, CCLM_{4M} further reduces this

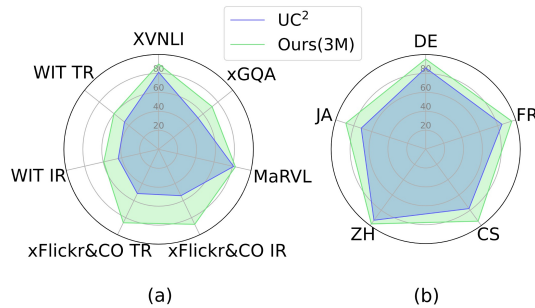


Figure 2: Visualization of cross-lingual transfer gap.

Methods	Flickr30K	MaRVL	xGQA	xFlickr&CO IR	TR
Ours	92.67	67.05	41.66	63.77	62.13
-w/o shared cross-attn (+5%)	92.49	66.67	36.76	63.73	62.01
-w/o shared FFN (+10%)	92.24	63.63	35.53	63.15	61.04
-w/ TLM	91.88	62.65	35.84	58.44	56.73
-w/ TLM + CL	92.34	65.00	36.13	63.42	61.33
-w/o parallel sentence pairs	91.90	58.37	28.80	44.11	43.24

Table 3: **Ablation study results.** Models w/o shared cross-attention and FFN are ablated variants where these modules are separately parameterized in the cross-lingual fusion model and the cross-modal fusion model. Models w/ TLM and TLM + CL are ablated variants where the multi-lingual objectives are that used in XLM-R and InfoXLM respectively, thus not unified with the multi-modal objectives. All compared models are pre-trained for 15 epochs.

gap to 4% and 2.8%. This further confirms that the proposed cross-view language modeling framework can effectively transfer multi-modal representations from English to other languages without language-specific fine-tuning.

In addition, we also visualize the multi-lingual text representations and image representations in CCLM and a baseline approach in Figure 3, which clearly shows our approach can better align multi-lingual image-text representations. It’s also noteworthy that the improved cross-lingual transfer ability does not sacrifice the model’s performance on English. In Table 8, we can see that CCLM performs competitively with state-of-the-art English VLMs on representative English vision-language tasks. For completeness, we also report per-language results on the IGLUE benchmark in Table 9, 10, 11, 12, and 13. Please refer to the Appendix for these analyses and results.

4.3 Ablation Study

We also conduct an in-depth ablation study to investigate the role of different design choices in the cross-view language modeling framework. We pre-train 5 ablated variants of CCLM where parallel sentence pairs, unified architecture, or unified objectives are ablated. All compared models are pre-trained with the same CC3M and WikiMatrix data (except that w/o parallel sentence pairs) for 15 epochs to ensure a fair comparison. The results are shown in Table 3. First, we find that separate parameterization of cross-attention and FFN modules in the cross-lingual and cross-modal fusion models leads to inferior results, especially for multi-lingual multi-modal understanding tasks such as xGQA. We also find that using common objectives in multi-lingual pre-training literature, which is different from the multi-modal objectives, underperforms the unified objectives used in our approach. These observations confirm the importance of unified architectures and objectives in the cross-view language modeling framework. Moreover, we find that the use of parallel sentence pairs also plays a very important role. This indicates that previous methods fail to fully exploit the potential of language pivoting for multi-lingual multi-modal pre-training.

5 Conclusion

In this paper, we introduce cross-view language modeling, a simple and effective framework that unifies cross-lingual and cross-modal pre-training. It considers cross-lingual and cross-modal pre-training as the same procedure of aligning the representation of two different views of the same object, thus using shared model architectures and training objectives for multi-lingual multi-modal pre-training. We train CCLM with the proposed framework and show that it advances the state-of-the-art on all downstream multi-lingual vision-language tasks by a large margin. More importantly, it surpasses the translate-test baseline for the first time, demonstrating the potential of multi-lingual multi-modal pre-training. We believe our model will become a foundation for future multi-lingual multi-modal research and serve as a strong baseline. Moreover, the cross-view language modeling framework also has the potential of unifying more modalities such as speech and video with the same architectures and objectives. We leave this for future work.

Acknowledgements

We would like to thank Hang Li, Jiase Chen, and Huiyun Yang at ByteDance for insightful comments in technical discussions. We also thank Yaoming Zhu at ByteDance for his generous assistance in data collection and valuable feedback.

References

- [1] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana, 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1202. URL <https://aclanthology.org/N18-1202>.
- [2] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. 2018.
- [3] 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, 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL <https://aclanthology.org/N19-1423>.
- [4] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *ArXiv preprint*, abs/1907.11692, 2019. URL <https://arxiv.org/abs/1907.11692>.
- [5] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.
- [6] Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. Unified language model pre-training for natural language understanding and generation. In *NeurIPS*, pages 13042–13054, 2019.
- [7] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *ArXiv preprint*, abs/1910.10683, 2019. URL <https://arxiv.org/abs/1910.10683>.
- [8] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.703. URL <https://aclanthology.org/2020.acl-main.703>.
- [9] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In Hugo Larochelle, Marc’Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL <https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfc4967418bfb8ac142f64a-Abstract.html>.

- [10] Zhiyi Fu, Wangchunshu Zhou, Jingjing Xu, Hao Zhou, and Lei Li. Contextual representation learning beyond masked language modeling. In *ACL (1)*, pages 2701–2714. Association for Computational Linguistics, 2022.
- [11] Wangchunshu Zhou, Dong-Ho Lee, Ravi Kiran Selvam, Seyeon Lee, and Xiang Ren. Pre-training text-to-text transformers for concept-centric common sense. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL <https://openreview.net/forum?id=3k20LaiHYL2>.
- [12] Wangchunshu Zhou, Tao Ge, Canwen Xu, Ke Xu, and Furu Wei. Improving sequence-to-sequence pre-training via sequence span rewriting. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 571–582, Online and Punta Cana, Dominican Republic, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.45. URL <https://aclanthology.org/2021.emnlp-main.45>.
- [13] Canwen Xu, Wangchunshu Zhou, Tao Ge, Furu Wei, and Ming Zhou. BERT-of-theseus: Compressing BERT by progressive module replacing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7859–7869, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.633. URL <https://aclanthology.org/2020.emnlp-main.633>.
- [14] Wangchunshu Zhou, Canwen Xu, Tao Ge, Julian J. McAuley, Ke Xu, and Furu Wei. BERT loses patience: Fast and robust inference with early exit. In Hugo Larochelle, Marc’Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL <https://proceedings.neurips.cc/paper/2020/hash/d4dd111a4fd973394238aca5c05bebe3-Abstract.html>.
- [15] Wangchunshu Zhou, Canwen Xu, and Julian McAuley. BERT learns to teach: Knowledge distillation with meta learning. In *ACL (1)*, pages 7037–7049. Association for Computational Linguistics, 2022.
- [16] Alexis Conneau and Guillaume Lample. Cross-lingual language model pretraining. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d’Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 7057–7067, 2019. URL <https://proceedings.neurips.cc/paper/2019/hash/c04c19c2c2474dbf5f7ac4372c5b9af1-Abstract.html>.
- [17] 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. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.747. URL <https://aclanthology.org/2020.acl-main.747>.
- [18] Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. MAD-X: An Adapter-Based Framework for Multi-Task Cross-Lingual Transfer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7654–7673, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.617. URL <https://aclanthology.org/2020.emnlp-main.617>.
- [19] Zewen Chi, Li Dong, Furu Wei, Nan Yang, Saksham Singhal, Wenhui Wang, Xia Song, Xian-Ling Mao, Heyan Huang, and Ming Zhou. InfoXLM: An information-theoretic framework for cross-lingual language model pre-training. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3576–3588, Online, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.280. URL <https://aclanthology.org/2021.naacl-main.280>.
- [20] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In Hanna M. Wallach, Hugo

- Larochelle, Alina Beygelzimer, Florence d’Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 13–23, 2019. URL <https://proceedings.neurips.cc/paper/2019/hash/c74d97b01eae257e44aa9d5bade97baf-Abstract.html>.
- [21] Hao Tan and Mohit Bansal. LXMERT: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5100–5111, Hong Kong, China, 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1514. URL <https://aclanthology.org/D19-1514>.
 - [22] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. VL-BERT: pre-training of generic visual-linguistic representations. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. URL <https://openreview.net/forum?id=SygXPaEYvH>.
 - [23] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *European conference on computer vision*, pages 104–120. Springer, 2020.
 - [24] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*, pages 121–137. Springer, 2020.
 - [25] Minheng Ni, Haoyang Huang, Lin Su, Edward Cui, Taroon Bharti, Lijuan Wang, Dongdong Zhang, and Nan Duan. M3p: Learning universal representations via multitask multilingual multimodal pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3977–3986, 2021.
 - [26] Mingyang Zhou, Luwei Zhou, Shuohang Wang, Yu Cheng, Linjie Li, Zhou Yu, and Jingjing Liu. Uc2: Universal cross-lingual cross-modal vision-and-language pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4155–4165, 2021.
 - [27] Emanuele Bugliarello, Fangyu Liu, Jonas Pfeiffer, Siva Reddy, Desmond Elliott, Edoardo Maria Ponti, and Ivan Vulić. Iglue: A benchmark for transfer learning across modalities, tasks, and languages. *ArXiv preprint*, abs/2201.11732, 2022. URL <https://arxiv.org/abs/2201.11732>.
 - [28] Haoyang Huang, Yaobo Liang, Nan Duan, Ming Gong, Linjun Shou, Daxin Jiang, and Ming Zhou. Unicoder: A universal language encoder by pre-training with multiple cross-lingual tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2485–2494, Hong Kong, China, 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1252. URL <https://aclanthology.org/D19-1252>.
 - [29] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for NLP. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 2790–2799. PMLR, 2019. URL <http://proceedings.mlr.press/v97/houlsby19a.html>.
 - [30] Yan Zeng, Xinsong Zhang, and Hang Li. Multi-grained vision language pre-training: Aligning texts with visual concepts. *ArXiv preprint*, abs/2111.08276, 2021. URL <https://arxiv.org/abs/2111.08276>.
 - [31] Wangchunshu Zhou, Yan Zeng, Shizhe Diao, and Xinsong Zhang. Vlua: A multi-task benchmark for evaluating vision-language models. *CoRR*, abs/2205.15237, 2022.

- [32] Aashi Jain, Mandy Guo, Krishna Srinivasan, Ting Chen, Sneha Kudugunta, Chao Jia, Yinfei Yang, and Jason Baldridge. Mural: multimodal, multitask retrieval across languages. *ArXiv preprint*, abs/2109.05125, 2021. URL <https://arxiv.org/abs/2109.05125>.
- [33] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL <https://openreview.net/forum?id=YicbFdNTTy>.
- [34] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in Neural Information Processing Systems*, 34, 2021.
- [35] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *ArXiv preprint*, abs/1807.03748, 2018. URL <https://arxiv.org/abs/1807.03748>.
- [36] Lingpeng Kong, Cyprien de Masson d’Autume, Lei Yu, Wang Ling, Zihang Dai, and Dani Yogatama. A mutual information maximization perspective of language representation learning. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. URL <https://openreview.net/forum?id=Syx79eBKwr>.
- [37] 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, 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1238. URL <https://aclanthology.org/P18-1238>.
- [38] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.
- [39] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73, 2017.
- [40] Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. Wiki-Matrix: Mining 135M parallel sentences in 1620 language pairs from Wikipedia. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1351–1361, Online, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.eacl-main.115. URL <https://aclanthology.org/2021.eacl-main.115>.
- [41] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 9992–10002. IEEE, 2021. doi: 10.1109/ICCV48922.2021.00986. URL <https://doi.org/10.1109/ICCV48922.2021.00986>.
- [42] 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. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.747. URL <https://aclanthology.org/2020.acl-main.747>.
- [43] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. URL <https://openreview.net/forum?id=Bkg6RiCqY7>.

- [44] 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. doi: 10.1162/tacl_a_00166. URL <https://aclanthology.org/Q14-1006>.
- [45] 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, 2016. Association for Computational Linguistics. doi: 10.18653/v1/W16-3210. URL <https://aclanthology.org/W16-3210>.
- [46] 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. *ArXiv preprint*, abs/1504.00325, 2015. URL <https://arxiv.org/abs/1504.00325>.
- [47] 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, 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-2066. URL <https://aclanthology.org/P17-2066>.
- [48] Xirong Li, Chaoxi Xu, Xiaoxu Wang, Weiyu Lan, Zhengxiong Jia, Gang Yang, and Jieping Xu. Coco-cn for cross-lingual image tagging, captioning, and retrieval. *IEEE Transactions on Multimedia*, 21(9):2347–2360, 2019.
- [49] Andrej Karpathy and Fei-Fei Li. Deep visual-semantic alignments for generating image descriptions. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 3128–3137. IEEE Computer Society, 2015. doi: 10.1109/CVPR.2015.7298932. URL <https://doi.org/10.1109/CVPR.2015.7298932>.
- [50] Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal, 2015. Association for Computational Linguistics. doi: 10.18653/v1/D15-1075. URL <https://aclanthology.org/D15-1075>.
- [51] Ning Xie, Farley Lai, Derek Doran, and Asim Kadav. Visual entailment: A novel task for fine-grained image understanding. *ArXiv preprint*, abs/1901.06706, 2019. URL <https://arxiv.org/abs/1901.06706>.
- [52] Željko Agić and Natalie Schluter. Baselines and test data for cross-lingual inference. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, 2018. European Language Resources Association (ELRA). URL <https://aclanthology.org/L18-1614>.
- [53] Jonas Pfeiffer, Gregor Geigle, Aishwarya Kamath, Jan-Martin O Steitz, Stefan Roth, Ivan Vulić, and Iryna Gurevych. xgqa: Cross-lingual visual question answering. *ArXiv preprint*, abs/2109.06082, 2021. URL <https://arxiv.org/abs/2109.06082>.
- [54] Drew A. Hudson and Christopher D. Manning. GQA: A new dataset for real-world visual reasoning and compositional question answering. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 6700–6709. Computer Vision Foundation / IEEE, 2019. doi: 10.1109/CVPR.2019.00686. URL http://openaccess.thecvf.com/content_CVPR_2019/html/Hudson_GQA_A_New_Dataset_for_Real-World_Visual_Reasoning_and_Compositional_CVPR_2019_paper.html.
- [55] Fangyu Liu, Emanuele Bugliarello, Edoardo Maria Ponti, Siva Reddy, Nigel Collier, and Desmond Elliott. Visually grounded reasoning across languages and cultures. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10467–10485, Online and Punta Cana, Dominican Republic, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.818. URL <https://aclanthology.org/2021.emnlp-main.818>.

- [56] Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. A corpus for reasoning about natural language grounded in photographs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6418–6428, Florence, Italy, 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1644. URL <https://aclanthology.org/P19-1644>.
- [57] Krishna Srinivasan, Karthik Raman, Jiecao Chen, Michael Bendersky, and Marc Najork. Wit: Wikipedia-based image text dataset for multimodal multilingual machine learning. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2443–2449, 2021.
- [58] Michał Ziemski, Marcin Junczys-Dowmunt, and Bruno Poulliquen. The united nations parallel corpus v1. 0. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC’16)*, pages 3530–3534, 2016.
- [59] Jörg Tiedemann. Parallel data, tools and interfaces in opus. In *Lrec*, volume 2012, pages 2214–2218. Citeseer, 2012.
- [60] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021.
- [61] Roy Schwartz, Jesse Dodge, Noah A Smith, and Oren Etzioni. Green ai. *Communications of the ACM*, 63(12):54–63, 2020.
- [62] Jingjing Xu, Wangchunshu Zhou, Zhiyi Fu, Hao Zhou, and Lei Li. A survey on green deep learning. *arXiv preprint arXiv:2111.05193*, 2021.
- [63] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

A Appendix

A.1 Limitations and Potential Social Impacts

In this paper, CCLM is pre-trained with CC3M and WikiMatrix as multi-modal and multi-lingual pre-training data, which are of moderate size. Also, we only experiment with base-size models. These choices are made because we want to make apple-to-apple comparison with previous works such as M³P and UC². However, there exists larger public available multi-lingual datasets (e.g., MultiUN [58], OPUS [59], etc.) and multi-modal datasets (e.g., CC12M, LAION [60], etc.). As suggested by the comparison between CCLM_{4M} and CCLM_{3M}, adding these large-scale datasets will probably lead to further performance improvements. Using larger models will also likely lead to some improvements. However, we do not run experiments with huge pre-training data and giant models due to environmental considerations [61, 62] and we try to make our experiments as “green” as possible.

As for social impact, multi-modal pre-trained models can be used in applications that help people with disability in one modality. Our work makes these applications applicable to minority people speaking non-English, and potentially low-resource languages. In sum, our work potentially enables deep learning technology to benefit more people, and is unlikely to have direct negative social impact.

A.2 Details of Pre-training Datasets

We pre-train CCLM on the combination of image-caption pairs and parallel texts. For image-caption pairs, we follow UC² and use their release translation-augmented version of CC3M dataset, which contains machine-translated captions in five different languages (German, French, Czech, Japanese, and Chinese). As for parallel text corpus, we collect a subset of the WikiMatrix [40] dataset containing parallel texts between English and other languages in the IGLUE benchmark. Table 4 shows the number of pairs per language. In total, the dataset consists of 19M parallel sentence pairs.

ES	FR	PT	RU	DE	VI	ID	AR	JA	ZH
3,130	2,645	2,322	1,598	1,467	998	974	968	841	783
EL	CS	TR	DA	BG	KO	BN	ET	TA	SW
609	509	455	412	353	281	269	241	61	51

Table 4: The number of parallel sentence pairs per language (K) in the subset of WikiMatrix.

A.3 Hyperparameter for Fine-Tuning

On IGLUE, we first finetune the model on English training set, and then evaluate zero-shot and few-shot performance on other languages. For both zero-shot and few-shot experiments, we use AdamW as our optimizer, with $\beta_1, \beta_2 = 0.9, 0.999$; weight decay is set to 0.01; learning rate scheduler is linear. For zero-shot experiments, we search hyperparameters for XVNLI and xFlickr&CO tasks. For both of them, we search learning rates $\{5e-6, 1e-5, 3e-5\}$. The final hyperparameters are shown in Table 5.

Task	XVNLI	xGQA	MaRVL	xFlickr&CO	WIT
Learning rate	1e-5	3e-5	3e-5	1e-5	3e-5
Batch size	128	256	80	128	128
Epochs	10	5	10	10	10
Max input length	80	40	40	80	80

Table 5: Hyperparameters used for IGLUE zero-shot finetuning.

For few-shot experiments, we search three learning rates $\{1e-6, 1e-5, 5e-5\}$. The few-shot data size and final hyperparameters are shown in Table 6. For all the tasks, we train the network for 60 epochs on each language, and evaluate every 10 epochs to search the best results.

Task	XVNLI	xGQA	MaRVL	xFlickr&CO
Shot number	48	48	20	100
Learning rate	1e-6	1e-6	1e-6	1e-6
Batch size	64	64	32	32
Epochs	60	60	60	60
Max input length	80	40	40	80

Table 6: Data size and hyperparameters used for IGLUE few-shot finetuning.

A.4 Zero-Shot on Retrieval

Table 7 reports results on multi-lingual image-text retrieval of CCLM without fine-tuning (zero-shot). We can observe that CCLM outperforms MURAL which is pre-trained on much larger data. Besides, the performance gap on non-English test sets of MURAL is larger, which shows our model has better cross-lingual transfer ability.

Model	Flickr30K				MSCOCO		
	EN	DE	FR	CS	EN	ZH	JA
MURAL _{base}	82.4	76.2	75.0	64.6	79.2	-	73.4
MURAL _{large}	89.2	83.5	83.1	77.0	84.4	-	81.3
CCLM _{3M}	83.7	79.1	76.7	73.9	81.5	79.5	76.8
CCLM _{4M}	89.5	85.2	84.9	84.1	91.7	89.9	88.1

Table 7: **Zero-shot results on multi-lingual image-text retrieval.** We compute the average Recall@K for both image-to-text retrieval and text-to-image retrieval with K = 1, 5, 10, as the evaluation metric. Results of compared models are directly copied from the corresponding papers.

A.5 Results on English Tasks

Table 8 reports CCLM performance on three representative English multi-modal tasks. We can observe that CCLM also has competitive performance compared to state-of-the-art English multi-modal baselines.

Methods	VQA2.0	NLVR2		MSCOCO(5K)	
	test-dev	dev	test-P	IR	TR
VinVL _{base}	75.95	82.05	83.08	58.10	74.60
ALBEF (4M)	74.54	80.24	80.50	56.80	73.10
VLMo (4M)	76.64	82.77	83.34	57.20	74.80
X-VLM (4M)	78.07	84.16	84.21	63.10	80.40
CCLM _{4M}	77.17	82.66	83.22	60.89	77.72

Table 8: **Results on representative English multi-modal tasks.** R@1 and accuracy are reported for MSCOCO (5K test set) and understanding tasks (VQA2.0 and NLVR2) respectively.

A.6 Visualization of Representations

Figure 3 visualizes several examples in xFlickr&CO test set in 2D space using t-SNE [63]. The image representations and text representations are the output [CLS] embeddings of the image encoder and the cross-lingual text encoder respectively. We can observe that CCLM’s text representations in different languages are more gathered and the distances between text representations and corresponding image representations are relatively shorter. This indicates our approach can better align multi-lingual image-text representations.

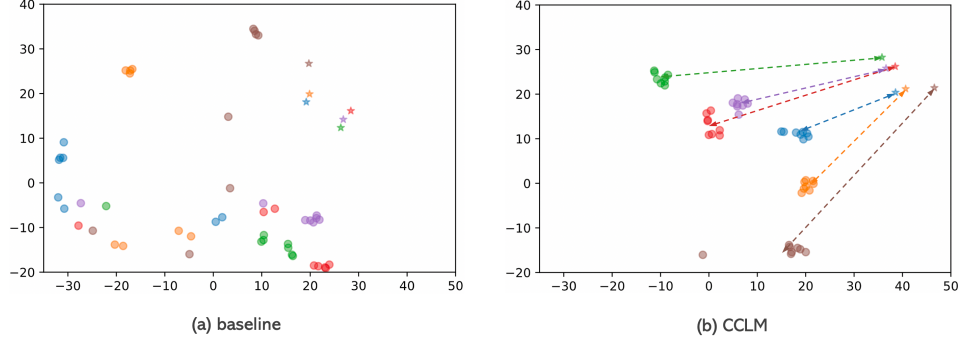


Figure 3: Visualization of image (denoted by stars) and text (denoted by points) representations. For a test example, there are eight texts in different languages. Points and stars in the same color are of the same test example. (a) is the ablated variant of CCLM that does not utilize parallel sentence pairs.

A.7 Per-Language Results

We also provide per-language results on the IGLUE benchmark in following tables, which supplement the ones provided in Table 1 in the main paper.

XVNLI Model	Language			
	ARB	SPA	FRA	RUS
<i>Zero-Shot</i>				
3M	71.04(0.98)	75.80(0.39)	78.14(0.33)	73.56(0.79)
4M	69.68(1.08)	73.65(1.45)	77.54(0.35)	72.40(0.33)
<i>Few-Shot</i>				
3M	71.82(0.10)	75.84(0.05)	78.62(0.05)	74.31(0.09)
4M	70.26(0.05)	75.06(0.05)	78.18(0.05)	73.27(0.30)

Table 9: **Full per-language results on XVNLI.** We use accuracy as evaluation metric for this task.

xGQA Model	Language						
	BEN	DEU	IND	KOR	POR	RUS	CMN
<i>Zero-Shot</i>							
3M	33.16(0.69)	51.62(0.21)	46.97(0.51)	39.80(0.70)	41.29(1.60)	36.51(1.44)	47.13(0.74)
4M	35.85(0.97)	56.68(0.20)	51.12(0.31)	41.14(0.59)	48.45(1.39)	39.83(0.55)	50.62(1.66)
<i>Few-Shot</i>							
3M	50.75(0.03)	54.60(0.02)	50.97(0.01)	46.85(0.03)	52.08(0.06)	49.44(0.05)	51.90(0.13)
4M	54.42(0.03)	59.25(0.09)	55.17(0.05)	49.33(0.11)	57.01(0.02)	53.91(0.07)	56.66(0.08)

Table 10: **Full per-language results on xGQA.** We use accuracy as evaluation metric for this task.

MaRVL Model	IND	SWA	Language TAM	TUR	CMN
<i>Zero-Shot</i>					
3M	67.58(1.05)	61.55(1.18)	60.28(0.65)	69.60(0.76)	70.52(0.30)
4M	71.66(1.56)	67.21(0.81)	60.36(1.73)	66.75(0.91)	69.86(1.01)
<i>Few-Shot</i>					
3M	69.86(0.32)	-	-	70.86(0.22)	70.87(0.60)
4M	73.35(0.05)	-	-	69.45(1.30)	74.90(0.26)

Table 11: **Full per-language results on MaRVL.** We use accuracy as evaluation metric for this task. MaRVL only has few-shot sets for IND, TUR and CMN.

xFlickr&CO Type	Model	DEU	SPA	IND	Language JPN	RUS	TUR	CMN
<i>Zero-Shot</i>								
IR	3M	67.73(0.59)	71.23(0.90)	62.38(0.17)	71.10(0.38)	72.83(0.20)	55.15(0.53)	71.05(0.48)
	4M	73.65(0.28)	79.62(0.46)	69.50(0.34)	75.68(0.40)	80.65(0.20)	65.08(0.38)	79.03(0.64)
TR	3M	66.88(0.48)	68.58(0.15)	60.33(0.30)	68.55(0.40)	69.90(0.13)	54.22(0.74)	69.12(0.14)
	4M	73.60(0.16)	78.38(0.35)	67.67(0.53)	73.45(0.31)	80.35(0.38)	63.22(0.21)	77.58(0.20)
<i>Few-Shot</i>								
IR	3M	66.50(0.10)	69.13(0.24)	61.18(0.13)	69.67(0.03)	70.70(0.13)	55.85(0.10)	69.25(0.05)
	4M	73.62(0.16)	79.02(0.08)	68.93(0.05)	77.35(0.13)	80.92(0.08)	65.45(0.09)	77.90(0.05)
TR	3M	67.70(0.16)	72.27(0.24)	63.26(0.08)	72.37(0.06)	73.17(0.18)	56.68(0.08)	71.58(0.06)
	4M	73.59(0.05)	79.80(0.10)	70.62(0.12)	78.7(0.09)	80.93(0.18)	68.03(0.08)	79.3(0.05)

Table 12: **Full per-language results on xFlickr&CO.** We take Recall@1 as evaluation metric for both image-to-text retrieval and text-to-image retrieval.

WIT Type	Model	ARB	BUL	DAN	Language ELL	EST	IND	JPN	KOR	TUR	VIE
<i>Zero-Shot</i>											
IR	3M	28.21(0.93)	26.28(0.81)	28.88(1.09)	32.57(1.92)	17.66(0.40)	32.89(0.46)	26.67(0.67)	18.01(0.63)	27.97(0.90)	35.45(0.16)
	4M	28.84(0.59)	24.96(0.83)	29.93(0.45)	32.98(1.15)	17.58(0.58)	33.89(0.72)	27.37(0.40)	17.29(0.77)	26.40(0.53)	35.55(0.06)
TR	3M	31.27(0.22)	26.51(0.94)	32.15(0.35)	32.96(0.59)	19.74(0.46)	33.53(0.27)	26.33(0.59)	19.01(0.60)	28.93(0.53)	36.13(0.41)
	4M	31.32(1.05)	26.18(0.37)	31.45(0.86)	32.22(0.70)	19.74(0.80)	34.35(0.24)	28.19(0.72)	19.57(0.68)	28.39(1.25)	36.13(0.70)

Table 13: **Full per-language results on WIT.** We take Recall@1 as evaluation metric for both image-to-text retrieval and text-to-image retrieval.