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

Yan Zeng[*†]
ByteDance

Wangchunshu Zhou[*]
ETH Zurich

Ao Luo[*]
Waseda University

Ziming Cheng[*]
Shanghai Jiao Tong University

Xinsong Zhang
ByteDance

Abstract

In this paper, we introduce Cross-View Language Modeling, a simple and effective pre-training framework that unifies cross-lingual and 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**, a **C**ross-lingual **C**ross-modal **L**anguage **M**odel, 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%. Moreover, CCLM is the first multi-lingual multi-modal pre-trained model that surpasses the translate-test performance of representative English vision-language models by zero-shot cross-lingual transfer.[1]

1 Introduction

Recently, the tremendous success of self-supervised language model pre-training (Peters et al., 2018; Radford et al., 2018; Devlin et al., 2019; Liu et al., 2019; Radford et al., 2019; Dong et al., 2019; Raffel et al., 2019; Lewis et al., 2020; Brown et al., 2020) has been expanded to the multi-lingual (Conneau and Lample, 2019; Conneau et al., 2020; Pfeiffer et al., 2020; Chi et al., 2021) and multi-modal (Lu et al., 2019; Tan and Bansal, 2019; Su et al., 2020; Chen et al., 2020; Li et al., 2020) 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, M^3P (Ni et al., 2021) and UC^2 (Zhou et al., 2021), 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 (Bugliarello et al., 2022) 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 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

[*]Equal Contribution. Work done at ByteDance.
[†]Correspondence to: zengyan.yanne@bytedance.com
[1]The code and pre-trained models are available at https://github.com/zengyan-97/CCLM.

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 5731–5746
July 9-14, 2023 ©2023 Association for Computational Linguistics

applicable in real-world applications. In contrast, multi-lingual pre-trained models such as XLM-R (Conneau et al., 2020) 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, M^3P 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 (Conneau et al., 2020; Chi et al., 2021). 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 over 10% and 30% on multi-lingual vision-language understanding and retrieval tasks in terms of accuracy and R@1 on IGLUE (Bugliarello et al., 2022), 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. Since previous work used different pre-training datasets, making direct comparison difficult, we also conduct an in-depth ablation study to investigate the contribution of different parts in our framework. The results show that use of parallel sentence pairs helps to fully exploit the potential of language pivoting for multi-lingual multi-modal pre-training and also confirm the importance of unified architectures and objectives in CCLM.

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)** CCLM advances the state-of-the-art of multi-lingual vision-language pre-training by a large margin. It also surpasses the translate-test baseline for the first time, demonstrating the potential of multi-lingual multi-modal pre-training. **(3)** We further scale up CCLM with massive pre-training data and larger model size. We will release our large-scale pre-trained multi-lingual multi-modal models to benefit a larger set of tasks and user groups and setup a strong and easily reproducible baseline for multi-lingual multi-modal research.

2 Related Work

Multi-lingual Pre-training Multilingual BERT (Devlin et al., 2019) 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 (Conneau and Lample, 2019), XLM-R (Conneau et al., 2020), and Unicoder (Huang et al., 2019) 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 (Pfeiffer et al., 2020) and InfoXLM (Chi et al., 2021) further improve multi-lingual pre-training via adapter (Houlsby et al., 2019) and contrastive learning.

Vision-Language Pre-training Inspired by the success of language model pre-training, a number of work (Lu et al., 2019; Tan and Bansal, 2019; Li et al., 2020; Chen et al., 2020; Zeng et al., 2021; Wang et al., 2022; Yu et al., 2022) 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 (Zhou et al., 2022). 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, M^3P (Ni et al., 2021) uses English as pivot and alternates between English-only vision-language pre-training and multi-lingual masked language modeling. UC^2 (Zhou et al., 2021), on the other hand, translates 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 (Jain et al., 2021) collects large-scale image-text pairs in 110 languages and pre-trains a dual encoder model via contrastive learning. MURAL achieves new state-of-the-art on multi-lingual image-text retrieval tasks. 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 (Dosovitskiy et al., 2021) first splits an image into non-overlapping patches, and then embeds these patches with transformer layers, yielding $\{\vec{v}_{\text{cls}}, \vec{v}_1, ..., \vec{v}_{N_1}\}$. 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 $\{\vec{w}_{\text{cls}}, \vec{w}_1, ..., \vec{w}_{N_2}\}$. 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 $\{\vec{x}_{\text{cls}}, \vec{x}_1, ..., \vec{x}_{N_2}\}$.

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 (Chi et al., 2021; Li et al., 2021). 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

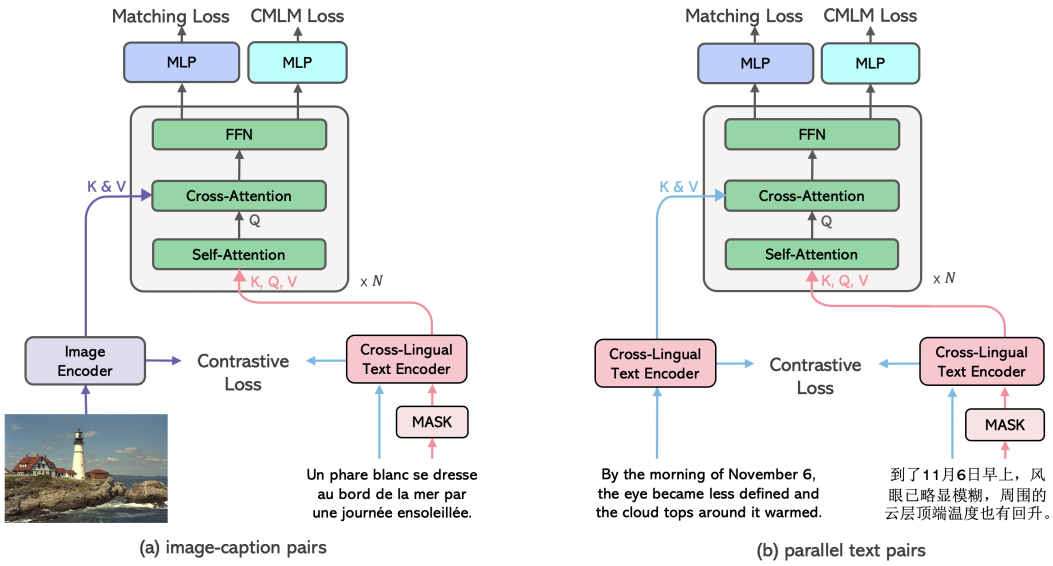

Figure 1: Illustration of the cross-view language modeling framework. CCLM takes two different views of the same object, i.e., either (a) image-caption pairs or (b) parallel sentence pairs, as input. CCLM first encodes the two views separately. Then the representations of the two views are fused by a Transformer-based 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.

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 (Oord et al., 2018) defined as:

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

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

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

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

where $|\tilde{A}| = |\tilde{B}| = N$ is the batch size, and we predict (a, b) pairs from in-batch negatives. $f_{\boldsymbol{\theta}}(a,b) = g_v(\vec{v}_{\text{cls}})^\top g_w(\vec{w}_{\text{cls}})/\tau$ given an image-caption pair or $f_{\boldsymbol{\theta}}(a,b) = g_w(\vec{w}^a_{\text{cls}})^\top g_w(\vec{w}^b_{\text{cls}})/\tau$ given a translation pair. $\vec{v}_{\text{cls}}$ and $\vec{w}_{\text{cls}}$ are the output [CLS] embedding of the image encoder [2] 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 $\mathcal{L}_{\text{nce}}$:

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

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, $f_{\boldsymbol{\theta}}(a,b) = \vec{v}^\top_{\text{true}}\vec{x}_{\text{cls}}(a,b)$, where $\vec{v}_{\text{true}}$ is a parametric vector.

The conditional MLM loss can also be interpreted as maximizing mutual information (Kong et al., 2020) 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:

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

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

where $f_{\boldsymbol{\theta}}(c, w_i) = \psi(w_i)^\top \vec{x}_i(\hat{a}, b)$. $\vec{x}_i$ 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: $\mathcal{L} = \mathcal{L}_{\text{cl}} + \mathcal{L}_{\text{match}} + \mathcal{L}_{\text{mlm}}$, 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

We pre-train CCLM on the combination of image-caption pairs and parallel multilingual texts. Appendix A.1 describes compared models in details.

Multi-modal Data For image-caption pairs, we follow the practice of UC^2 to make a fair comparison and use their released translation-augmented version of CC3M dataset. It contains the original CC3M image-caption pairs (Sharma et al., 2018) and machine-translated captions in five different languages (German, French, Czech, Japanese, and Chinese). This multi-modal dataset is widely utilized by previous work, including UC^2, mUNITER and xUNITER. We denote this variant as CCLM^{3M}. In additional to this setting, we leverage large-scale vision language pre-training by utilizing the pre-trained weights of X^2-VLM (Zeng, 2021; Zeng et al., 2022) which has been trained on more than 1B image-text pairs in English. Based on it we apply the proposed framework for multi-lingual multi-modal pre-training.

Multi-lingual Data Previous work such as mUNITER, xUNITER, and M^{3}P use large-scale monolingual texts in different languages, namely multilingual Wikipedia 101G dataset, for multilingual alignment. Differently, we propose to utilize parallel text corpus. We collect a subset of the WikiMatrix (Schwenk et al., 2021) dataset containing parallel texts between English and other languages in the IGLUE benchmark. Appendix A.2 shows the number of pairs per language. In total, the dataset consists of 19M parallel sentence pairs.

4.1.2 Implementation Details

CCLM$_{\text{base}}$ consists of 12 Transformer layers for the image encoder and the text encoder respectively. CCLM$_{\text{large}}$ consists of 24 layers for each encoder. The fusion encoder contains 6 Transformer layers for both CCLM$_{\text{base}}$ ($d = 768$) and CCLM$_{\text{large}}$ ($d = 1024$). In total, CCLM$_{\text{base}}$ and CCLM$_{\text{large}}$ consist of $\sim$ 420M and $\sim$ 970M parameters respectively. Following existing models such as M^{3}P and UC^2, we also utilize XLM-R (Conneau et al., 2020) as the text encoder. Concretely, CCLM^{3M} is initialized with a pre-trained image encoder (Liu et al., 2021b) and XLM-R. CCLM is initialized with the pre-trained X^2-VLM (Zeng, 2021; Zeng et al., 2022) and XLM-R.

In pre-training, the image encoder takes images of resolution of 224×224 as input for pre-training. During fine-tuning, we increase the image resolution to 384×384 and interpolate the positional embeddings of image patches following Dosovitskiy et al. (2021). The maximum sequence length is set to 30 and 64 for image captions and parallel multilingual texts respectively. We apply mixed precision for pre-training. We use the AdamW (Loshchilov and Hutter, 2019) optimizer with a weight decay of 0.02. We mix different types of data in a training batch. Following UC^2, to make a fair comparison, we train CCLM^{3M} for 30 epochs on 8 NVIDIA A100 GPUs and the batch size is set to 1024, which tasks $\sim$ 1.5 days. The learning rate is warmed-up to $1e^{-4}$ in the first 2500 steps and decayed linearly. We train CCLM$_{\text{base}}$ and CCLM$_{\text{large}}$ for 40 epochs.

4.1.3 Downstream Tasks

We evaluate CCLM on the IGLUE benchmark (Bugliarello et al., 2022), a recently released benchmark for evaluating multi-lingual multi-modal pre-training, and a multi-lingual image-text retrieval benchmark including the multi-lingual version of Flickr30K (Young et al., 2014; Elliott et al., 2016) and MSCOCO (Chen et al., 2015). Note that CCLM can also be applied on generation tasks such as image captioning by following the adaptation strategy of X-VLM (Zeng et al., 2022; Zeng and Nie, 2021).

XVNLI: The Cross-lingual Visual NLI dataset is collected by combining SNLI (Bowman et al., 2015) with its multi-modal (Xie et al., 2019) and multi-lingual (Agić and Schluter, 2018) 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 (Pfeiffer et al., 2021) is collected by manually translating the GQA (Hudson and Man-

Model	NLI	QA	Reasoning	Retrieval			
	XVNLI	xGQA	MaRVL	xFlickr&CO		WIT	
				IR	TR	IR	TR
Translate everything to English and use English-only model (Translate-Test)							
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
Fine-tune model on English training set (Zero-Shot)							
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^3P	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
$CCLM^{3M}_{base}$	74.64	42.36	65.91	67.35	65.37	27.46	28.66
$CCLM_{base}$	74.78	48.12	68.49	76.94	76.22	33.90	35.26
$CCLM_{large}$	**78.95**	**56.25**	**74.83**	**83.78**	**83.46**	**43.74**	**44.88**

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. In the zero-shot setting, the models are fine-tuned on English train sets and directly evaluated on target languages. We report few-shot results in Appendix A.4.

ning, 2019) 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. (2021).

MaRVL: The Multicultural Reasoning over Vision and Language dataset (Liu et al., 2021a) 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 (Suhr et al., 2019) 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 crowd-source image descriptions in 6 other languages. Similarly, the Wikipedia-based Image Text dataset (Srinivasan et al., 2021) is collected from Wikipedia in 108 languages. We follow the data preprocessing and splitting details in IGLUE for both datasets.

Multi30K: This dataset (Elliott et al., 2016) extended Flickr30K (Young et al., 2014) 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 (Chen et al., 2015) by translating the captions into Japanese (Yoshikawa et al., 2017) and Chinese (Li et al., 2019). 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 (2015). As for Chinese, we use the COCO-CN split (Li et al., 2019).

For all retrieval tasks, we follow previous work (Li et al., 2021) and X-VLM (Zeng et al., 2021). During fine-tuning, we optimize $\mathcal{L}_{\text{cl}}$ and $\mathcal{L}_{\text{match}}$. For inference, we first compute similarity for all images and texts, and then take the top-k candidates and calculate the final ranking scores using the fusion model.

4.2 Experimental Results

4.2.1 Results on IGLUE Benchmark

Table 1 shows CCLM performance on the IGLUE benchmark. First, for zero-shot cross-lingual transfer, we can see that $CCLM^{3M}_{base}$ outperforms all compared models by a substantial margin while pre-trained on the same multi-modal data. Specifically, compared to UC^2, the prior state-of-the-art, $CCLM^{3M}_{base}$ 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

Model	Multi30K				MSCOCO		
	EN	DE	FR	CS	EN	ZH	JA
M^3P	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
$CCLM^{3M}_{base}$	95.3	92.4	92.1	91.2	93.1	92.2	93.2
$CCLM_{base}$	97.2	94.6	95.5	94.8	95.4	93.2	95.7
$CCLM_{large}$	**97.8**	**95.8**	**96.6**	**96.2**	**95.6**	**94.0**	**96.1**

Table 2: **Results on multi-lingual image-text retrieval** in all-language fine-tune setting, where a model is fine-tuned on the combination of training data in all languages. Following previous work, 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. We additionally report results in other fine-tune settings in Appendix A.5.

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 language modeling framework can better align multi-lingual multi-modal representations with unified objectives.

We also find that the performance of our framework can be significantly improved by leveraging large-scale image-text pre-training in English ($CCLM_{base}$) and/or scaling up the model size ($CCLM_{large}$). Notably, CCLM is the first multi-lingual multi-modal pre-trained model that substantially outperforms the translate-test results of representative English VLMs tested in the IGLUE benchmark. 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.

4.2.2 Results on Multi-lingual Retrieval

Table 2 gives the results on the multi-lingual image-text retrieval benchmark. When pre-trained on the same multi-modal data, $CCLM^{3M}_{base}$ substantially outperforms UC^2, the prior state-of-the-art, with an averaged improvement of over 10% (in terms of averaged recall) across four languages on Multi30K. This confirms that our approach can better align multi-lingual multi-modal representations. $CCLM^{3M}_{base}$ even outperforms MURAL. 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 without fine-tuning in Appendix A.5.

We also find that the cross-view language modeling framework yields better performance if leveraging large-scale pre-training on image-text pairs in English ($CCLM_{base}$) and/or scaling up the model size ($CCLM_{large}$), which is consistent with the experimental results on the IGLUE benchmark. It confirms that the proposed framework is scalable to both massive data and larger model size.

4.2.3 Cross-lingual Transfer Gap

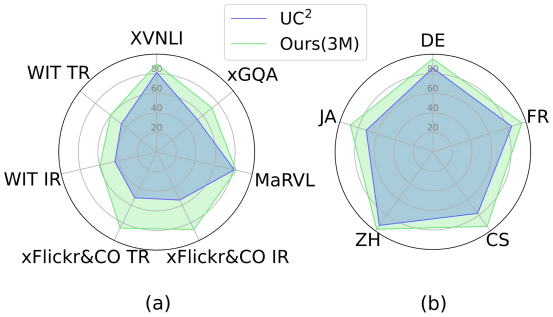

Figure 2: Visualization of cross-lingual transfer gap.

In addition to absolute cross-lingual transfer results reported in Table 1 and Table 2, 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 chat 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 UC^2 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 Appendix A.5, we can see that for M^3P, the absolute zero-shot cross-lingual transfer gap between EN-CS and EN-JA in Multi30K 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 UC^2 is reduced to 13.2% and 16.4%, demonstrating the effectiveness of using machine-translated captions for multi-lingual multi-modal pre-training. $CCLM^{3M}_{base}$ further reduces this gap to 5.4% and 4.4%. This confirms that the proposed cross-view language modeling

Methods	Multi30K	MaRVL	xGQA	xFlickr&CO	
				IR	TR
Ours	**92.67**	**67.05**	**41.66**	**63.77**	**62.13**
-w/o shared cross-attn	92.49	66.67	36.76	63.73	62.01
-w/o shared FFN	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 variants where the multi-lingual objectives are that used in XLM-R and InfoXLM, which are not unified with the multi-modal objectives. All compared models are pre-trained for 15 epochs.

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 Appendix A.6, which clearly shows our approach can better align multi-lingual image-text representations.

4.3 Ablation Study

Since previous work such as M^3P, UC^2, and MURAL all use different pre-training datasets, making direct comparison difficult, we conduct an in-depth ablation study to investigate the contribution 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 the use of parallel sentence pairs 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. On the other hand, CCLM variant trained without parallel sentences in Table 3 which uses the same pre-training dataset as UC^2 still significantly outperforms previous models such as M^3P and UC^2.

We then compare other ablated variants which all utilized parallel sentence pairs. We find that separate parameterization of cross-attention and FFN modules for the cross-lingual and the cross-modal task in the fusion model leads to inferior results, especially for multi-lingual multi-modal understanding tasks such as xGQA.

Moreover, we conduct ablation study on loss functions. We mainly consider multi-lingual objectives because the multi-modal objective combination of itc+mlm+itm is the de-facto choice for multi-modal loss (Li et al., 2021; Zeng et al., 2021). We find that using common objectives in the multi-lingual pre-training literature underperforms our unified objective. These observations confirm the importance of unifying architectures and objectives 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. Moreover, it surpasses the translate-test baseline for the first time, demonstrating the potential of multi-lingual multi-modal pre-training. Furthermore, the experimental results also confirm that the proposed framework is scalable to massive data and larger model sizes. 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 audio and video with the same architectures and objectives. We leave this for future work.

Limitations

In this paper, we pre-train CCLM with moderate multi-modal data, e.g. CC3M, to make a fair comparison with previous work such as M^3P and UC^2. We leverage large-scale vision language pre-training simply by utilizing the pre-trained weights of X^2-VLM which has been pre-trained on billion-scale image-text pairs in English. Collecting more image-text pairs in different languages will very likely lead to further performance improvements. Moreover, there exists larger public available multi-lingual datasets, such as MultiUN (Ziemski et al., 2016) and OPUS (Tiedemann, 2012). Leveraging more multi-lingual datasets for pre-training should also yield a more powerful multi-lingual multi-modal model.

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.

Acknowledgements

We would like to thank Hang Li, Jiaze 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

Željko Agić and Natalie Schluter. 2018. Baselines and test data for cross-lingual inference. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. 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. Association for Computational Linguistics.

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. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Emanuele Bugliarello, Fangyu Liu, Jonas Pfeiffer, Siva Reddy, Desmond Elliott, Edoardo Maria Ponti, and Ivan Vulić. 2022. Iglue: A benchmark for transfer learning across modalities, tasks, and languages. *ArXiv preprint*, abs/2201.11732.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. 2015. Microsoft coco captions: Data collection and evaluation server. *ArXiv preprint*, abs/1504.00325.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. Uniter: Universal image-text representation learning. In *European conference on computer vision*, pages 104–120. Springer.

Zewen Chi, Li Dong, Furu Wei, Nan Yang, Saksham Singhal, Wenhui Wang, Xia Song, Xian-Ling Mao, Heyan Huang, and Ming Zhou. 2021. 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. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *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.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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. Association for Computational Linguistics.

Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. 2019. Unified language model pre-training for natural language understanding and generation. In *NeurIPS*, pages 13042–13054.

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. 2021. 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.

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

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for NLP. In *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.

Haoyang Huang, Yaobo Liang, Nan Duan, Ming Gong, Linjun Shou, Daxin Jiang, and Ming Zhou. 2019. 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. Association for Computational Linguistics.

Drew A. Hudson and Christopher D. Manning. 2019. 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.

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

Andrej Karpathy and Fei-Fei Li. 2015. 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.

Lingpeng Kong, Cyprien de Masson d'Autume, Lei Yu, Wang Ling, Zihang Dai, and Dani Yogatama. 2020. 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.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. 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. Association for Computational Linguistics.

Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. 2021. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in Neural Information Processing Systems*, 34.

Xirong Li, Chaoxi Xu, Xiaoxu Wang, Weiyu Lan, Zhengxiong Jia, Gang Yang, and Jieping Xu. 2019. Coco-cn for cross-lingual image tagging, captioning, and retrieval. *IEEE Transactions on Multimedia*, 21(9):2347–2360.

Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. 2020. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*, pages 121–137. Springer.

Fangyu Liu, Emanuele Bugliarello, Edoardo Maria Ponti, Siva Reddy, Nigel Collier, and Desmond Elliott. 2021a. 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. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *ArXiv preprint*, abs/1907.11692.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021b. 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.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

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

Minheng Ni, Haoyang Huang, Lin Su, Edward Cui, Taroon Bharti, Lijuan Wang, Dongdong Zhang, and Nan Duan. 2021. M3p: Learning universal representations via multitask multilingual multimodal pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3977–3986.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *ArXiv preprint*, abs/1807.03748.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. 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. Association for Computational Linguistics.

Jonas Pfeiffer, Gregor Geigle, Aishwarya Kamath, Jan-Martin O Steitz, Stefan Roth, Ivan Vulić, and Iryna Gurevych. 2021. xgqa: Cross-lingual visual question answering. *ArXiv preprint*, abs/2109.06082.

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. 2020. 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. Association for Computational Linguistics.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *ArXiv preprint*, abs/1910.10683.

Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. 2021. WikiMatrix: 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. Association for Computational Linguistics.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. 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. Association for Computational Linguistics.

Krishna Srinivasan, Karthik Raman, Jiecao Chen, Michael Bendersky, and Marc Najork. 2021. 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.

Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2020. 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.

Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. 2019. 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. Association for Computational Linguistics.

Hao Tan and Mohit Bansal. 2019. 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. Association for Computational Linguistics.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in opus. In *Lrec*, volume 2012, pages 2214–2218. Citeseer.

Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(11).

Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. 2022. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *International Conference on Machine Learning*, pages 23318–23340. PMLR.

Ning Xie, Farley Lai, Derek Doran, and Asim Kadav. 2019. Visual entailment: A novel task for fine-grained image understanding. *ArXiv preprint*, abs/1901.06706.

Yuya Yoshikawa, Yutaro Shigeto, and Akikazu Takeuchi. 2017. 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. Association for Computational Linguistics.

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

Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. 2022. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*.

Danting Zeng. 2021. Multi task learning based framework for multimodal classification. In *Proceedings of the Third Workshop on Multimodal Artificial Intelligence*, pages 30–35, Mexico City, Mexico. Association for Computational Linguistics.

Yan Zeng and Jian-Yun Nie. 2021. An investigation of suitability of pre-trained language models for dialogue generation – avoiding discrepancies. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4481–4494, Online. Association for Computational Linguistics.

Yan Zeng, Xinsong Zhang, and Hang Li. 2021. Multi-grained vision language pre-training: Aligning texts with visual concepts. *ArXiv preprint*, abs/2111.08276.

Yan Zeng, Xinsong Zhang, Hang Li, Jiawei Wang, Jipeng Zhang, and Wangchunshu Zhou. 2022. X ^{2}-vlm: All-in-one pre-trained model for vision-language tasks. *arXiv preprint arXiv:2211.12402*.

Mingyang Zhou, Luowei Zhou, Shuohang Wang, Yu Cheng, Linjie Li, Zhou Yu, and Jingjing Liu. 2021. Uc2: Universal cross-lingual cross-modal vision-and-language pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4155–4165.

Wangchunshu Zhou, Yan Zeng, Shizhe Diao, and Xinsong Zhang. 2022. Vlue: A multi-task benchmark for evaluating vision-language models. *CoRR*, abs/2205.15237.

Michał Ziemski, Marcin Junczys-Dowmunt, and Bruno Pouliquen. 2016. The united nations parallel corpus v1. 0. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3530–3534.

A Appendix

A.1 Compared Models

mUNITER and xUNITER: A multi-lingual variant of the UNITER (Chen et al., 2020) model pre-trained by Liu et al. (2021a). 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.

M^{3}P: 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.

UC^2: 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.

A.2 Details for Multi-lingual Data

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 Results on English Tasks

Table 5 reports CCLM performance that is pre-trained on COCO, VG, SBU, and CC3M, on three common English multi-modal tasks. We can observe that CCLM also has very competitive performance compared to strong English multi-modal baselines.

Methods	**VQA2.0**	**NLVR2**		**MSCOCO**(5K)	
	test-dev	dev	test-P	IR	TR
VinVL$_{\text{base}}$	75.95	82.05	83.08	58.10	74.60
ALBEF (4M)	74.54	80.24	80.50	56.80	73.10
CCLM$^{\text{4M}}_{\text{base}}$	77.17	82.66	83.22	60.89	77.72

Table 5: **Results on common English multi-modal tasks.** R@1 and Accuracy are reported for MSCOCO (5K test set) and understanding tasks respectively.

A.4 Few-Shot Results on IGLUE

Table 6 gives results on IGLUE benchmark. 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. 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. We omit few-shot evaluation on the WIT dataset because this setup is also omitted in IGLUE. We find that similar to existing models, CCLM can also benefit from few-shot learning with a few examples in the target languages.

A.5 More Results on Retrieval Tasks

Table 7 reports results on multi-lingual image-text retrieval of CCLM. 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.

We also report multi-lingual image-text retrieval results without fine-tuning (zero-shot) in Table 8. M^{3}P and UC^2 do not report their zero-shot retrieval performances. We can observe that CCLM$^{\text{3M}}_{\text{base}}$ outperforms MURAL which is pre-trained on much larger data. Besides, the performance gap on non-

Model	NLI	QA	Reasoning	Retrieval			
	XVNLI	xGQA	MaRVL	xFlickr&CO		WIT	
				IR	TR	IR	TR
Few-shot train English fine-tuned model on target languages (Few-Shot)							
mUNITER	53.95	37.21	53.41	8.54	9.32	-	-
xUNITER	60.55	40.68	57.46	14.30	13.54	-	-
M^3P	59.36	41.04	49.79	13.21	12.26	-	-
UC^2	63.68	42.95	58.32	19.79	17.59	-	-
$CCLM^{3M}_{base}$	**75.15(.03)**	**50.94(.02)**	**70.53(.18)**	**66.04(.05)**	**68.15(.04)**	-	-

Table 6: **Few-Shot 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 model, mean and standard deviation (in brackets) of 3 different runs with different random seeds are reported.

Model	Multi30K				MSCOCO		
	EN	DE	FR	CS	EN	ZH	JA
English-only Fine-tune (Zero-Shot)							
M^3P	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
$CCLM^{3M}_{base}$	**94.8(.11)**	**90.3(.08)**	**90.9(.38)**	**89.4(.21)**	**93.2(.05)**	**91.0(.18)**	**88.8(.06)**
Single-Language Fine-tune							
M^3P	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
$CCLM^{3M}_{base}$	**94.8(.11)**	**91.9(.16)**	**90.6(.18)**	**88.9(.05)**	**93.2(.05)**	**90.2(.24)**	**93.3(.26)**

Table 7: **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 model, mean and standard deviation (in brackets) of 3 different runs with different random seeds are reported.

Model	Multi30K				MSCOCO		
	EN	DE	FR	CS	EN	ZH	JA
$MURAL_{base}$	82.4	76.2	75.0	64.6	79.2	-	73.4
$CCLM^{3M}_{base}$	**83.7**	**79.1**	**76.7**	**73.9**	**81.5**	**79.5**	**76.8**

Table 8: **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.

English test sets of MURAL is larger, which shows our model has better cross-lingual transfer ability.

A.6 Visualization of Representations

Figure 3 visualizes several examples in xFlickr&CO test set in 2D space using t-SNE (Van der Maaten and Hinton, 2008). The image representations and text representations are the output [CLS] embeddings of the image encoder and the cross-lingual text encoder respectively.

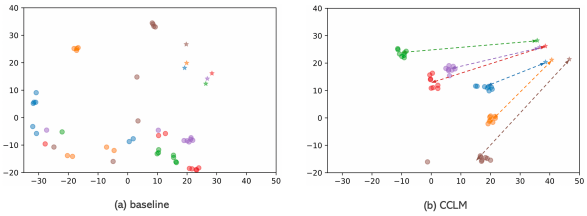

Figure 3: Visualization of image (denoted by stars) and text (denoted by points) representations. For a test example, there are eight texts in different languages. Points and stars in the same color are of the same test example. (a) is the ablated variant of CCLM that does not utilize parallel sentence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

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.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
section 6

☑ A2. Did you discuss any potential risks of your work?
section 6

☑ A3. Do the abstract and introduction summarize the paper's main claims?
section 1

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

experiment section

☑ B1. Did you cite the creators of artifacts you used?
experiment section

☒ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
they are commonly used datasets

☒ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
they are commonly used datasets

☒ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
they are commonly used datasets

☒ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
they are commonly used datasets

☒ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
they are commonly used datasets

C ☑ Did you run computational experiments?

section 4

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
section 4.1.2

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
section 4.1.1, 4.1.2, and 4.1.3; we report our best hyperparameter values with code release.

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Appendix Table 5 and Table 6. For our model, mean and standard deviation (in brackets) of 3 different runs with different random seeds are reported.

☐ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Not applicable. Left blank.

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

☐ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
No response.

☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
No response.

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
No response.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No response.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
No response.