Machine Intelligence Research
www.mi-research.net

20(2), April 2023, 220-232
DOI: 10.1007/s11633-022-1414-4

Multimodal Pretraining from Monolingual to Multilingual

Liang Zhang Ludan Ruan Anwen Hu Qin Jin

School of Information, Renmin University of China, Beijing 100872, China

Abstract: Multimodal pretraining has made convincing achievements in various downstream tasks in recent years. However, since the majority of the existing works construct models based on English, their applications are limited by language. In this work, we address this issue by developing models with multimodal and multilingual capabilities. We explore two types of methods to extend multimodal pretraining model from monolingual to multilingual. Specifically, we propose a pretraining-based model named multilingual multimodal pretraining (MLMM), and two generalization-based models named multilingual CLIP (M-CLIP) and multilingual acquisition (MLA). In addition, we further extend the generalization-based models to incorporate the audio modality and develop the multilingual CLIP for vision, language, and audio (CLIP4VLA). Our models achieve state-of-the-art performances on multilingual vision-text retrieval, visual question answering, and image captioning benchmarks. Based on the experimental results, we discuss the pros and cons of the two types of models and their potential practical applications.

Keywords: Multilingual pretraining, multimodal pretraining, cross-lingual transfer, multilingual generation, cross-modal retrieval.

Citation: L. Zhang, L. Ruan, A. Hu, Q. Jin. Multimodal pretraining from monolingual to multilingual. *Machine Intelligence Research*, vol.20, no.2, pp.220–232, 2023. http://doi.org/10.1007/s11633-022-1414-4

1 Introduction

Vision, audio, and language modalities are all important information carriers in our daily life. Thus, the development of multimodal understanding models has attracted much research interest[1–4]. Among them, multimodal pretraining achieves great success in various vision-language tasks. However, for the text modality, existing methods are typically based on monolinguals such as English. Since people from all over the world communicate in different languages, the application of the current models is limited by monolingualism.

In this paper, we focus on developing models that handle different modalities and languages. We explore the following two types of methods to achieve both multimodal and multilingual capability: 1) pretraining-based methods that design weakly-supervised objectives based on large-scale multilingual multimodal data and 2) generalization-based methods that extend existing monolingual models into multilingual models.

For the pretraining-based method, we propose multilingual multimodal pretraining (MLMM) on a large scale multilingual image-text corpus. To encourage MLMM to build the alignment between the image and multilingual text at both coarse-grained and fine-grained levels, we adopt four popular pretraining objectives named image-text matching (ITM), masked language modeling (MLM), masked region feature regression (MRFR), and masked region classification (MRC), respectively. Experimental results show that MLMM performs well on cross-lingual transfer, and achieves state-of-the-art performance on multilingual image-text retrieval, visual question answering, and image captioning tasks. We further demonstrate that fine-tuning MLMM with languages belonging to the same language family can bring improvements.

For the generalization-based method, we extend the monolingual multimodal model CLIP[4] into multilingual via two strategies: multilingual knowledge distill (MKD)[5] and multilingual acquisition (MLA)[6]. MKD[5] introduces a multilingual encoder to process non-English sentences. The multilingual encoder is trained through knowledge distillation[7] from the monolingual text encoder; MLA[6] extends the monolingual text encoder by inserting language acquirers to handle other languages. We demonstrate that these two generalization-based methods achieve comparable performance with the pretraining-based methods, while requiring much fewer computing resources and data resources of each language. However, they typically require translation pairs between source and target languages and show worse cross-lingual transfer ability. On the basis of the experimental results, we discuss the pros and cons of the three methods to achieve multilingual and multimodal capability.

Additionally, as audio is a commonly used modality in human daily life, we further propose multilingual CLIP for vision, language, and audio (CLIP4VLA) that endows the two generalization-based models with the ability of audio processing. Experimental results show that multilingual CLIP4VLA outperforms previous state-of-the-art methods on multilingual video-text retrieval

Research Article
Special Issue on Large-scale Pre-training: Data, Models, and Fine-tuning
Manuscript received July 7, 2022; accepted December 28, 2022
Recommended by Associate Editor Cheng-Lin Liu

benchmarks. We also point out that introducing audio modality can indeed bring significant improvements towards video-text retrieval, which is consistent with the fact that audio is semantically related and complementary to visual and language modalities in video.

The contributions of this paper can be summarized as follows:

1) We propose MLMM, a multilingual and multimodal pretraining model. It achieves competitive results on multilingual image-text retrieval, multilingual visual question answering and multilingual image captioning tasks.

2) We adopt the generalization-based methods MKD and MLA to extend the monolingual multimodal model CLIP, which consume much less training data and computing resources, and achieve promising results. We further propose CLIP4VLA to extend on audio modality, which achieves competitive performance on multilingual video-text retrieval.

3) We analyze the two types of methods to achieve multilingual and multimodal, and provide some practical suggestions based on the experiments. We hope these analyses could promote further study on multilingual and multimodal understanding.

2 Related work

Multilingual multimodal pretraining

Many efforts have been made to develop multilingual and multimodal pretraining models. The main challenge of these approaches is the shortage of non-English image-text data. Multitask multilingual multimodal pre-training (M3P)[8] assembles English image-text pairs and multilingual sentences for pretraining. In this way, the model can establish the alignment between non-English sentences and images by pivoting on English. M3P further proposes code-switched training objectives to enhance the learning of cross-lingual representations conditioned on the same image. Unsupervised cross-lingual cross-modal pre-training (UC2)[9] obtains multilingual image-text data through machine translation, and introduces additional pretraining objectives for fine-grained text-image alignment. Multimodal multitask retrieval across languages (MURAL)[10] collects large-scale multilingual image-text pairs and bilingual translation pairs from the Web[11] and pretrains with image-text and text-text contrastive objectives. We use a similar strategy with UC2 to construct the pretraining dataset by translating the Chinese dataset RUC-CAS-WenLan[12] and the English dataset Conceptual Captions[13, 14] into seven languages. Different from UC2, which adopts fine-grained objectives to all image-text pairs equally, we observe that some image-text pairs crawled from the web are weakly correlated[12, 15]. Thus, we only adopt coarse-grained objectives on weakly correlated pairs. For strongly correlated pairs, we adopt both coarse-grained and fine-grained objectives.

Generalize monolingual to multilingual

In contrast to training with multilingual corpus from scratch, some works explore generalizing the existing monolingual models into multilingual models. This can be achieved by knowledge transfer techniques such as knowledge distillation[7] and adapters[16]. Multilingual knowledge distillation[5] distills the task-specific knowledge from the monolingual model to a multilingual model using translation pairs. Multiple adapters for cross-lingual transfer (MAD-X)[17] extends monolingual models to low resource languages through adapters[16]. It trains the adapters with the MLM objective on monolingual data[17]. MLA[6] handles other languages by inserting language acquirers into the monolingual model. Since task specific knowledge has been learned from monolingual data, these generalization-based methods typically consume less multilingual data than pretraining-based methods.

3 Methodology and experiments

In this section, we introduce our proposed multimodal and multilingual models: the pretraining based model MLMM, the generalization based models MKD and MLA, and the vision-language-audio pretrained model CLIP4VLA.

3.1 Multilingual vision-language pretraining

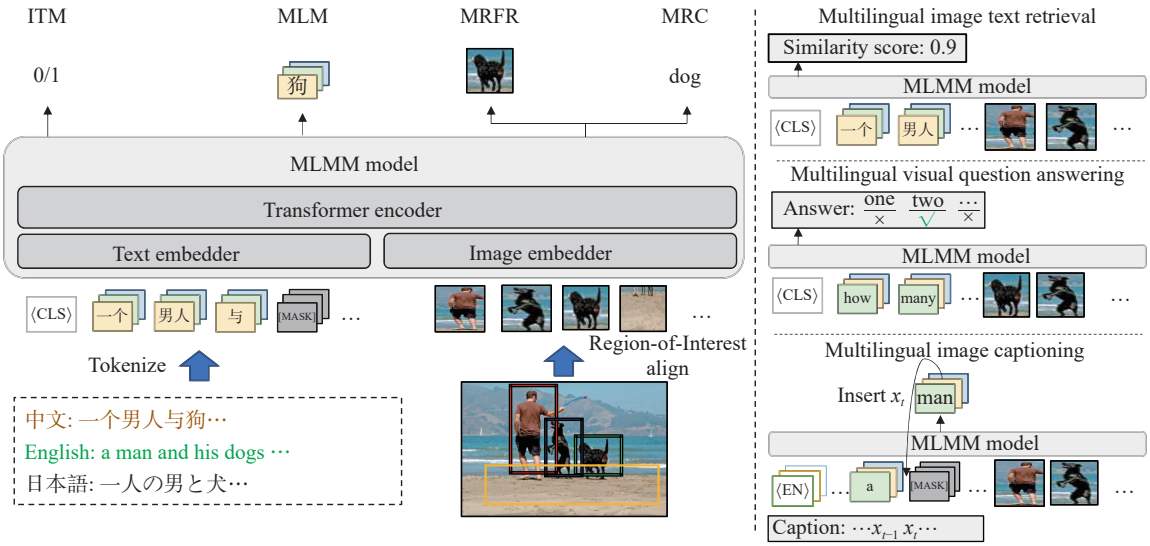
We propose MLMM pretraining to learn universal representations across different languages and modalities.

3.1.1 Architecture

As shown in Fig. 1(a), MLMM is a BERT-like model that is pretrained with multilingual image-text pairs. Given multilingual image-text pairs (I, T^*), where T^* can be in different languages, we first process I and T^* with the image embedder and text embedder, respectively. The image embedder extracts region-of-interest features of image I via bottom-up-attention detector[18]. Following UNITER[3], we add location embeddings for each region and perform linear projection[19] to keep the dimension consistent. We denote the final region embeddings as $\boldsymbol{R} = \{r_1, \cdots, r_k\}$. The text embedder first tokenizes T^* into subwords via SentencePiece[20], and then embeds the subwords with a shared multilingual look-up table. After adding learnable position embedding[21], the text is represented as a sequence of word embeddings $\boldsymbol{W} = \{w_1, \cdots, w_n\}$. With concatenated multimodal sequence $[\boldsymbol{R}; \boldsymbol{W}]$ as the input, the multilingual multimodal transformer learns cross-lingual and cross-modal representations. We do not restrict the behavior of the self-attention in the transformer so that the tokens from the two modalities can attend to each other freely. This enables fine-grained interaction between images and captions.

3.1.2 Pretraining objectives

Different downstream tasks require different granularity of interaction between images and texts. We design four pretraining objectives to learn the alignment between



(a) MLMM structure and pre-training objectives (b) Downstream fine-tuning

Fig. 1 Overview of MLMM

image and multilingual text from coarse-grained to fine-grained:

Image-text matching (ITM)

In ITM, the model is trained to discriminate whether the content of the input image and text is matched[3]. We randomly sample 50% negative image-text pairs by replacing the image or text with another sample from the whole dataset D. Label $y \in \{0, 1\}$ denotes whether the sample pair is matched or not. We choose the output feature corresponding to the ⟨CLS⟩ token to represent the whole pair, and pass it to a linear projection layer with sigmoid activation to calculate the matching score (denoted as f). The loss function of ITM is shown in (1)

$$\mathcal{L}_{\text{ITM}} = -\mathbb{E}_{(I,T^*)\sim D}[y \log(f(I, T^*; \theta)) + (1-y)\log(1 - f(I, T^*; \theta))] \quad (1)$$

where $\mathbb{E}$ denotes the expectation in all formulas, $y = 1$ if (I, T^*) is a positive pair, otherwise $y = 0$, θ is the parameter of MLMM. As we sample negative pairs from the whole dataset, the image context and text are usually very different in negative pairs. Thus, ITM encourages MLMM to learn coarse-grained alignment between image and text.

Masked language modeling (MLM)

Since ITM can handle alignment at the instance level, we also perform MLM[3], which recovers the masked phrases at the token level. In MLM, we follow the same masking strategy described in BERT[21]. Formally, denoting the masked tokens and unmasked tokens as $\boldsymbol{W}_m$ and $\boldsymbol{W}_{\setminus m}$ respectively, the loss function of MLM is defined as follows:

$$\mathcal{L}_{\text{MLM}} = -\mathbb{E}_{(\boldsymbol{R},\boldsymbol{W})\sim D}[\log(P(\boldsymbol{W}_m | \boldsymbol{R}, \boldsymbol{W}_{\setminus m}; \theta))]. \quad (2)$$

Unlike BERT[21], which uses text-only context to predict masked tokens, MLMM can recover the masked tokens with image regions. This encourages MLMM to learn fine-grained alignment between image regions and text tokens.

Masked region feature regression (MRFR)

Since MLM encourages fine-grained alignment from the perspective of language modality, we also perform a pretraining objective that learns from the perspective of vision modality. MRFR requires the MLMM to recover the masked region features conditioned on multimodal context[3]. Similar to MLM, we also randomly select 15% regions to perform masking. The input embeddings of masked regions are replaced by zero vectors. Suppose $\boldsymbol{R}_m$ and $\boldsymbol{R}_{\setminus m}$ are masked and unmasked regions respectively, the loss function of MFRF is defined as

$$\mathcal{L}_{\text{MRFR}} = \mathbb{E}_{(\boldsymbol{R},\boldsymbol{W})\sim D} \|\boldsymbol{R}_m - F(\boldsymbol{R}_m | \boldsymbol{R}_{\setminus m}, \boldsymbol{W}; \theta)\|_2 \quad (3)$$

where F denotes the output feature of MLMM.

Masked region classification (MRC)

Apart from recovering the masked region features, MRC uses the pseudo category of each region predicted by the detector[3]. This objective promotes MLMM to identify the semantic categories of the masked regions. Since the pseudo category from the detector can be noisy, we use the distribution of categories produced by the detector as a soft label $\tilde{C}$, and optimize Kullback-Leibler divergence (KLD) to make the predicted distribution $\hat{C}$ close to $\tilde{C}$:

$$\mathcal{L}_{\text{MRC}} = \mathbb{E}_{(\boldsymbol{R},\boldsymbol{W})\sim D} \text{KLD}(\tilde{C}(\boldsymbol{R}_m) \| \hat{C}(\boldsymbol{R}_m | \boldsymbol{R}_{\setminus m}, \boldsymbol{W}; \theta)). \quad (4)$$

3.1.3 Pretraining corpus of MLMM

We pretrain MLMM on three large-scale image-text corpora: RUC-CAS-WenLan[12], Conceptual Captions (CC3M)[13], and Conceptual-12M (CC12M)[14].

RUC-CAS-WenLan (RCW)[12] is an image-text dataset crawled from Chinese Web sources such as news, Baidu Baike, and Weibo. It contains 30 million image-text pairs, and we randomly select 15 million image-text pairs from it to build the multilingual corpus. We use NiuTrans Cloud Platform[1] to automatically translate the Chinese text in the corpus into 6 target languages including English, German, French, Czech, Japanese and Korean. Unlike the traditional image captioning dataset, the image and text in RUC-CAS-WenLan are weakly correlated[12]. As shown in Fig. 2(a), there is little fine-grained correspondence between words and image regions. Thus, fine-grained pretraining objectives such as MLM, MRC, and MRFR are not suitable for RUC-CAS-WenLan. We therefore use coarse-grained ITM objective for this dataset.

CC3M and CC12M[13, 14] are image-text datasets collected from English webpages. Compared with RUC-CAS-WenLan, these two datasets adopt a more restricted filtering strategy. In addition to filtering out low-quality and illegal images and texts separately, they also apply image-based and text-based filtering to make sure the image content can be described by the corresponding text[13]. As shown in Fig. 2(b), the image and text are strongly correlated. Thus, building fine-grained correspondence between the two modalities is possible. Therefore, we optimize MLMM with all objectives described in Section 3.1.2 on CC3M and CC12M. We also translate the English texts into 6 target languages (Chinese, German, French, Japanese, and Korean). We denote the ensemble of CC3M[13] and CC12M[14] as CC15M.

3.1.4 Implementation details

The multilingual multimodal transformer adopts the $\text{BERT}_{\text{large}}$ setting, which consists of 24 transformer layers with 1 024-d hidden state. We initialize all parameters of the multilingual multimodal transformer and text embedder with XLM-R[22]. The visual embedder is implemented as bottom-up-attention pretrained on VisualGenome[23] with the ResNet-101[24] backbone. Regions with confidence scores higher than 0.2 are input to the model. Following UNITER[3], at each training step, we sample mini-batches from one dataset, and optimize one objective from the four objectives described in Section 3.1.2. Each mini-batch contains different languages of image-text pairs. We use 40 NVIDIA Geforce A100 with 40 GB memory to train MLMM. The total batch size is 60. We pretrain MLMM for 600 000 steps, and the total pretraining costs 15 days to converge. We use a linear scheduler during training. The learning rate increases from 0 to 1×10^{-4} during the first 5% of training, and then decays to 0 in the remaining steps.

3.1.5 Experiments

We conduct experiments to evaluate MLMM on three downstream tasks: multilingual image-text retrieval, visual question answering and image captioning.

[1] https://niutrans.com

Multilingual image-text retrieval

In this task, models are asked to find the most semantically similar image given texts in different languages or vice versa. As shown in Fig. 1(b), we use the feature corresponding to the ⟨CLS⟩ token to calculate the similarity score between the image and text. To comprehensively evaluate the cross-lingual transfer ability and language capacity of the multilingual models, we conduct experiments under the three settings as follows[22]:

1) Zero-shot: The models are directly tested on the downstream datasets without fine-tuning. This setting evaluates the generalization ability of models to unseen images and texts.

2) Finetune on English: The models are finetuned with English downstream image-text pairs, and tested on other languages. This setting measures the cross-lingual generalization ability of models to unseen languages.

3) Finetune on all: The models are jointly finetuned with downstream data in all languages. This setting evaluates the language capacity of the models.

We evaluate multilingual image-text retrieval on two downstream datasets: Multi30K[25] and MSCOCO[26–28]. Multi30K provides 31 K images. Each image has 5 captions in English and German, and 1 caption in French and Czech. MSCOCO[26–28] contains 123 K images with 5 captions in English and Japanese. We follow the train/val/test splits defined in [25] and [27] for Multi30K and MSCOCO, respectively. We report the average score of Recall@1, Recall@5, and Recall@10 for each dataset and language following previous works[8–10]. The comparison between MLMM and state-of-the-art models is shown in Table 1. MLMM significantly outperforms existing models under the zero-shot setting in all languages. We consider the reason to be that our pretraining dataset is more suitable for image-text retrieval since each text in all languages is paired with an image. Therefore, the alignment between the image and each language can be directly modeled through paired data. In contrast, as the pretraining dataset of M3P and MURAL only contains English image-text pairs, the relationship between image and other languages is hard to build. Although the pretraining dataset of UC2 has image-text pairs for each language, the amount is much smaller than ours. When finetuned on English, MLMM also achieves better cross-lingual transfer compared with other models. When finetuned on all languages, MLMM achieves comparable results with the state-of-the-art model MURAL which was pretrained with much larger datasets. It reveals that it is not easy to learn cross-lingual and cross-modal representations since a large training dataset is essential to achieve convincing performance.

Cross-lingual transfer experiments

We further conduct experiments to evaluate the generalization ability of MLMM across languages. We finetune MLMM on each language, and then directly evaluate the model on other languages. Note that the number

中文：周末哥哥放学回家饭堂大叔忙碌起来我和饭堂大叔已习惯在周三就开始计划周末做点什么好吃的[偷笑][偷笑]。

English: At the weekend, my brother came home from school and uncle canteen was busy. Uncle canteen and I were used to planning to do something delicious on Wednesday

中文：企业团队投资创业者交易讨论分析图股市交易，股票图概念。

English: Business team investment entrepreneur trading discussing and analysis graph stock market trading, stock chart concept

(a) RUC-CAS-WenLan

English: Woman caresses the dog in winter snowy cold day

中文：女人在冬天雪天雪地里爱抚狗

English: Young girls playing in the garden with a bicycle

中文: 年轻女孩们在花园里玩自行车

(b) Conceptual captions

Fig. 2 Image-text samples from (a) RUC-CAS-WenLan and (b) Conceptual Captions.

Table 1 Multilingual image-text retrieval results on Multi30K and MSCOCO. Trans: Translate-train; FT-En: Finetune on English; FT-All: Finetune on all. ‡: These models are jointly finetuned on COCO-CN[29], while others are not. The best results are in bold, and the second best are underlined.

	Method	Training data	# Images	# Text	# Params.	Multi30K				MSCOCO	
						EN	DE	FR	CS	EN	JA
Zero-shot	M3P[8]	CC3M+Wiki	3.3M	101B	566M	57.9	36.8	27.1	20.4	63.1	33.3
	UC2[9]	Trans(CC3M)	3.3M	19.6M	478M	66.6	62.5	60.4	55.1	70.9	62.3
	MURAL[10]	Trans(CC12M)+EOBT	12M	512M	300M	80.9	76.0	75.7	68.2	78.1	72.5
	MURAL[10]	AT+MBT	1.8B	7.8B	300M	82.4	76.2	75.0	64.6	79.2	73.4
	MLMM	Trans(CC15M+RCW)	30M	210M	1B	**86.3**	**81.7**	**81.2**	**79.3**	**83.6**	**81.1**
FT-En	M3P[8]	CC3M+Wiki	3.3M	101B	566M	87.4	82.1	67.3	65.0	88.6	56.0
	UC2[9]	Trans(CC3M)	3.3M	19.6M	478M	87.2	83.8	77.6	74.2	88.1	71.7
	MLMM	Trans(CC15M+RCW)	30M	210M	1B	**89.7**	**84.4**	**84.2**	**81.7**	**89.8**	**84.2**
FT-all	M3P‡[8]	CC3M+Wiki	3.3M	101B	566M	87.7	82.7	73.9	72.2	88.7‡	87.9‡
	UC2‡[9]	Trans(CC3M)	3.3M	19.6M	478M	88.2	84.5	83.9	81.2	88.1‡	87.5‡
	MURAL[10]	Trans(CC12M)+EOBT	12M	512M	300M	91.0	87.3	86.4	82.4	89.4	87.4
	MURAL[10]	AT+MBT	1.8B	7.8B	300M	**92.2**	**88.6**	**87.6**	84.2	88.6	88.4
	MLMM	Trans(CC15M+RCW)	30M	210M	1B	90.7	87.2	86.0	**85.3**	**89.8**	**90.1**

of texts is inconsistent between languages in Multi30K. This can lead to unfair comparison across languages, as the lack of data in the finetune language can lead to poor cross-lingual transferability to other languages. To address this issue, we construct a new dataset M30K$_{\mathrm{T}}$ based on Multi30K[25] by translating each English sentence into other 6 languages. As a result, all languages share the same amount of image-text pairs. Table 2 shows the cross-lingual transfer results on M30K$_{\mathrm{T}}$. We observe that the performance generalized from other languages can be very close (with $\Delta_{\max} \leq 1.9\%$ average recall) to the finetuning performance. This suggests that MLMM can generalize the image-text knowledge learned from each language to other languages. Additionally, we observe that two languages in different language families often produce worse transfer results. For example, the model finetuned with German (Indo-European language) performs worse on Chinese (Sino-Tibetan language) and Korean (language isolate), and the model finetuned on Korean performs worse on Indo-European languages including German, French, and Czech. We consider that the discrepancy between different language families makes cross-lingual transfer harder.

Influence of language families during joint finetuning

Previous works[8, 9] have shown that joint fine-tuning with multiple languages can benefit the image-text retrieval for each language. However, the influence of the language family is unexplored. The above cross-lingual transfer experiments have shown that knowledge transfer

Table 2 Cross-lingual transfer experiments on $M30K_T$. The best result of each language is bolded, and the worst result is marked red. Δ_{max} is the max performance gap between fine-tuning and transferring from other languages.

	EN	DE	FR	CS	ZH	JA	KO
EN	**91.9**	89.1	88.6	87.5	89.6	88.3	88.2
DE	90.7	**89.9**	88.6	87.8	89.2	88.5	88.1
FR	90.7	89.3	**90.1**	88.2	89.5	88.3	88.4
CS	90.6	89.0	88.5	**89.3**	89.8	88.9	88.8
ZH	90.8	89.3	89.0	88.0	**90.7**	88.8	88.6
JA	90.9	89.3	89.1	88.3	89.6	**90.2**	89.1
KO	90.7	88.6	88.3	87.6	89.5	88.9	**90.0**
Δ_{max}	1.3	1.3	1.8	1.7	1.5	1.9	1.9

between different language families is more difficult. Therefore, we conduct experiments to verify whether fine-tuning on the same/different language families benefits the target language. We still evaluate the models on the $M30K_T$. We divide the 7 languages into two groups by whether they are Indo-European languages (English, French, German, Czech) or Non-Indo-European languages (Chinese, Korean, Japanese). Models are fine-tuned on the following four settings: 1) Each language; 2) Non-Indo-European languages; 3) Indo-European languages; 4) All languages. We investigate the performance of the Indo-European languages. The results are shown in Table 3.

Table 3 Finetune MLMM on $M30K_T$ with different language family

Finetune setting	Test language			
	EN	DE	FR	CS
Each	91.9	89.9	90.1	89.3
Non-Indo-European	91.3	89.9	89.7	89.0
Indo-European	**91.9**	**90.5**	**90.3**	**89.7**
All	91.4	89.8	89.7	89.3

Comparing to fine-tuning on each language, jointly fine-tuning on four Indo-European languages brings improvements, and fine-tuning on other languages does not help. This suggests that knowledge sharing is easier between languages in the same language family. If there is a shortage in the target language, we can leverage the data in the same language family for fine-tuning to achieve better performance.

Multilingual visual question answering

We evaluate MLMM on the visual question answering task to verify its fine-grained understanding ability. We finetune and evaluate MLMM on the English dataset VQA2.0[30] and Japanese dataset VQA-VG-JA[31], respectively. As shown in Fig. 1(b), we treat VQA as a recognition task, where the model selects the correct answer from the answer pool of the whole dataset. We select the top-3 129 and top-3 000 frequent answers as the answer pool of VQA2.0 and VQA VG Japanese respectively[9]. Table 4 shows the accuracy comparison between MLMM and other pretraining based models. We can see that MLMM outperforms UC2 significantly on both datasets. This indicates that MLMM learns cross-lingual and cross-modal knowledge, which not only benefits coarse-grained tasks such as image-text retrieval, but is also applicable to more challenging tasks such as VQA.

Table 4 Visual question answering in English (VQA2.0) and Japanese (VQA VG JA). Metric: Accuracy (%).

Model	VQA2.0 test-dev	VQA VG JA
PCATT[31]	–	19.2
$UNITER_{CC}$[3]	71.22	22.7
UC2[9]	71.48	34.2
MLMM	**73.21**	**35.4**

Multilingual image captioning

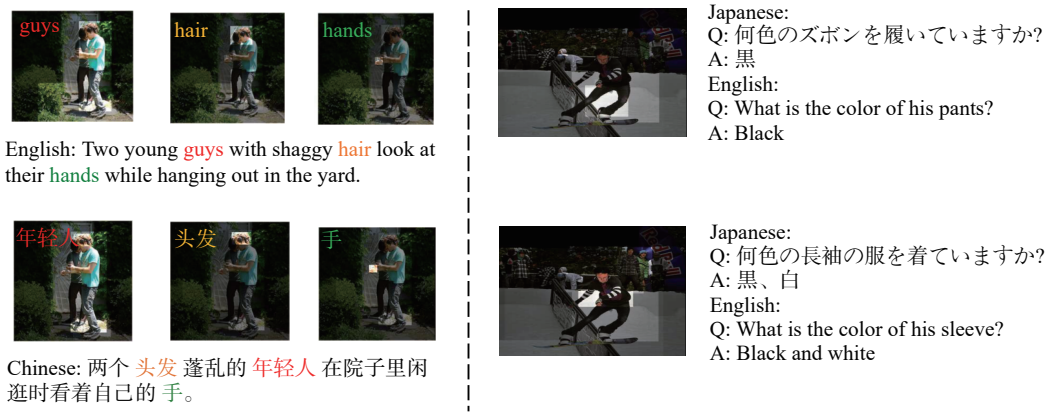
We further evaluate the language generation ability of MLMM through the multilingual image captioning task. Following VLP[3], we finetune MLMM to recover masked tokens by previous tokens. During inference, MLMM generates tokens in a caption autoregressively through mask prediction as shown in Fig. 1(b). MLMM uses a language token to identify the language to generate[32, 33]. We evaluate multilingual image captioning on Multi30K[25] and use BLEU4 (B@4)[34] and CIDEr (C)[35] as metrics. The experimental results in Table 5 show that MLMM outperforms the two pretraining based models BertGen[36] and M3P[8] in all languages. This indicates that MLMM has strong multilingual generation ability in addition to understanding. Additionally, we find fine-tuning on multiple languages outperforms separately fine-tuning on each language. This may be because descriptions in different languages contain diverse details of the image. This reveals the necessity of multilingual, since learning to describe an image in different languages can promote the model to capture comprehensive vision features.

3.1.6 Visualization

We visualize the token-level and sentence-level attention distributions over image regions. As shown in

Table 5 Multilingual image captioning on Multi30K

Model	EN	DE	FR	CS
	B@4/C	B@4/C	B@4/C	B@4/C
BertGen (Finetune on EN+DE)[36]	27.0/58.7	17.8/34.2	–/–	–/–
M3P (Finetune on each)[8]	26.1/57.2	16.1/43.8	7.5/36.1	4.0/28.5
M3P (Finetune on all)[8]	26.5/59.4	16.6/44.3	8.7/40.1	**5.4**/31.1
MLMM (Finetune on each)	23.7/54.8	18.8/55.0	6.4/50.7	2.1/30.4
MLMM (Finetune on all)	**28.8/66.0**	**20.3/57.4**	**8.7/72.2**	4.3/**48.1**



(a) Token-level attention (b) Senetence-level attention

Fig. 3 Visualization of attention distribution over image regions. Region is brighter if it has a higher attention score.

Fig. 3(a), tokens in different languages with the same semantic meaning attend to similar image regions. In Fig. 3(b), MLMM can attend to the related regions regardless of whether the question is in English or Japanese. This indicates that MLMM successfully models the correspondence across different modalities and languages. It is also consistent with the fact that all human beings share a similar vision system[37].

3.2 Multilingual vision-language generalization

Although MLMM achieves convincing performance, it costs huge computing resources and requires considerable training data on various languages, as it learns cross-lingual and cross-modal representations simultaneously from scratch. In this section, we consider extending a monolingual multimodal pretraining model to multilingual one in order to achieve multilingual and multimodal capability with much less effort.

3.2.1 Multilingual CLIP

Multilingual CLIP (M-CLIP)[38] generalizes multimodal pretraining into multilingual through MKD[5]. MKD[5] is first proposed to extend monolingual texts into multilingual for NLP tasks such as bilingual sentence retrieval. It treats the pretrained monolingual text encoder as a fixed teacher model and distills knowledge to a student multilingual encoder with translation pairs. Concretely, supposing there is a translation pair (S, T), where S is written in the source language (English) and T is written in the target language, the multilingual encoder $\mathcal{M}$ is expected to produce a representation of T that is close to the representation of S encoded by text encoder Ψ. This is achieved by minimizing L2 distance between the two representations as follows:

$$\mathcal{L}_{\text{MKD}} = \mathbb{E}_{(S,T)\in D}\|\Psi(S,\theta_\Psi) - \mathcal{M}(T,\theta_\mathcal{M})\|_2 \quad (5)$$

where θ_Ψ is frozen during training, and $\theta_\mathcal{M}$ is trainable.

3.2.2 Multilingual acquisition

MLA[6] empowers a multimodal pretraining model with multilingual capability by inserting lightweight language acquirers into the pretrained text encoder. The language acquirers are implemented as bottleneck MLPs. They are inserted after all transformer layers of the pretrained text encoder. The training procedure of MLA consists of two stages: the native language transfer (NLT) stage and the language exposure (LE) stage. Specifically, we denote the pretrained text encoder inserted with language acquirers as language acquisition encoder Ψ'. In the NLT stage, given target sentence T, Ψ' is optimized to match the source feature produced by the text encoder Ψ given the source sentence S. The loss function can be written as follows:

$$\mathcal{L}_{\text{NLT}} = \mathbb{E}_{(S,T)\in D}\|\Psi(S,\theta_\Psi) - \Psi'(T,\theta_\Psi,\theta_{\text{LA}},\theta_{\text{emb}})\|_2 \quad (6)$$

where θ_{LA} represents the parameters of language

acquirers, θ_{emb} is the multilingual embedding matrix. Note that each target language has a unique group of language acquirers θ_{LA}, but all target languages share the same θ_{emb}. During training, θ_{Ψ} is kept fixed. Compared with the multilingual encoder in M-CLIP, the language acquisition encoder shares most of the parameters with the pretrained text encoder. Thus the number of trainable parameters of MLA is less than that of M-CLIP.

In the LE stage, the language acquisition encoder is supposed to project the sentences in the target language into a shared multimodal space[39]. Given image-text pairs $\{(v_i, t_i)\}_{i=1}^{N}$, where $T = \{t_i\}_{i=1}^{N}$ are in target languages, the loss function of LE is defined as follows:

$$\mathcal{L}_{\text{LE}} = \frac{1}{2}(\text{NCE}(\Phi(V), \Psi'(T)) + \text{NCE}(\Psi'(T), \Phi(V))) \tag{7}$$

$$\text{NCE}(X, Y) = \frac{1}{B}\sum_{i=1}^{B} \log \frac{\exp(\text{sim}(x_i, y_i))}{\sum_{k=1}^{B} \exp(\text{sim}(x_i, y_i))} \tag{8}$$

where Φ is the image encoder of CLIP[4], whose parameters are also kept fixed during training, $\text{sim}(\cdot,\cdot)$ is the cosine similarity of two vectors.

3.2.3 Implementation details

We randomly sample 300 K image-text pairs (denoted as CC300K) from CC3M to train both M-CLIP and MLA. Note that both methods generalize the ViT-B-32 version of CLIP[4] into multilingual models. For M-CLIP, we initialize the multilingual encoder with bert-base-multilingual-cased[21]. Each mini-batch may contain translation pairs between English and different languages. The batch size is 128, and the learning rate is set to 5×10^{-5}. The model is trained for 117 500 steps. For MLA, we sample mini-batches of one language for each step, and iteratively train over all languages. We also initialize the multilingual embedding with bert-base-multilingual-cased[21], but randomly initialize the language acquirers. The NLT stage is trained for 117 500 steps with a learning rate of 1×10^{-4}. The LE stage is trained for 11 750 steps with a learning rate of 3×10^{-6}. The batch size is 128 in both stages.

3.2.4 Experiments

We evaluate both M-CLIP and MLA on multilingual image-text retrieval following the three settings described in Section 3.1.5. The comparison between generalization-based and pretraining-based models is shown in Table 6.

In general, M-CLIP and MLA achieve comparable performance with state-of-the-art pretraining-based models, while with much fewer computing resources and parameters. Specifically, since MLA handles other languages through lightweight acquirers and maintains the English performance of CLIP[4], it performs the best on English under both Finetune-on-English and Finetune-on-all settings. In contrast, M-CLIP performs better on other languages because it introduces a heavier multilingual encoder to handle other languages. After fine-tuning on English, we observe that both generalization methods transfer well to French. The reason is that French shares many similar words with English. Since CLIP[4] performs well on English, the generalization-based methods can easily transfer this advantage to French. Additionally, both M-CLIP and MLA perform worse than MLMM under the zero-shot setting. We consider the reason to be that MLMM benefits from large-scale pretraining and learns general representations across various domains.

3.2.5 Discussions

Based on the experimental results in Section 3.1.5 and Section 3.2.4, we discuss the pros and cons of MLMM, M-CLIP and MLA methods in Table 7, from which we can provide some recommendations for method selection in

Table 6 Multilingual image-text retrieval results on Multi30K and MSCOCO. PT: Pretraining-based methods; GL: Generalization-based methods; #TP: Number of trainable parameters. The best results are in bold, and the second best are underlined.

	Method	Training data	Type	Costs (GPU·days)	#TP	Multi30K				MSCOCO	
						EN	DE	FR	CS	EN	JA
Zero-shot	MURAL[10]	Trans(CC12M)+EOBT	PT	512 TPU d	300 M	80.9	76.0	75.7	68.2	78.1	72.5
	MLMM	Trans(CC15M+RCW)	PT	600 A100 d	1 B	**86.3**	**81.7**	**81.2**	**79.3**	**83.6**	**81.1**
	M-CLIP	Trans(CC300K)	GL	0.5 V100 d	178 M	82.1	77.1	75.2	72.3	78.5	73.6
	MLA_{CLIP}	Trans(CC300K)	GL	0.5 V100 d	108 M	84.4	78.7	77.7	70.8	79.4	74.9
FT-EN	MLMM	Trans(CC15M+RCW)	PT	600 A100 d	1 B	89.7	**84.4**	84.2	81.7	**89.8**	**84.2**
	M-CLIP	Trans(CC300K)	GL	0.5 V100 d	178 M	90.3	83.5	**85.2**	**82.1**	88.5	82.8
	MLA_{CLIP}	Trans(CC300K)	GL	0.5 V100 d	108 M	**92.0**	82.6	85.1	76.2	89.3	80.4
FT-all	MURAL[10]	Trans(CC12M)+EOBT	PT	512 TPU d	300 M	91.0	87.3	86.4	82.4	89.4	87.4
	MLMM	Trans(CC15M+RCW)	PT	600 A100 d	1 B	90.7	87.2	86.0	85.3	**89.8**	**90.1**
	M-CLIP	Trans(CC300K)	GL	0.5 V100 d	178 M	90.6	**88.2**	**87.8**	**85.6**	88.9	88.9
	MLA_{CLIP}	Trans(CC300K)	GL	0.5 V100 d	108 M	**92.0**	86.8	85.4	82.3	89.3	88.1

Table 7 Pros and cons of pretraining-based and generalization-based models

Method	Pros	Cons
MLMM	Strong cross-lingual transfer and zero-shot ability. Handles multiple languages in one model. Free from translation pairs	Large data consumption, computing costs, and model size.
M-CLIP	Small data consumption, computing costs, and model size. Strong performance on target languages. Handles multiple languages in one model.	Limited to the performance of the monolingual model. Performs worse on English. Requires translation pairs.
MLA	Least data consumption, computing costs, and model size. Preforms well on English.	Limited to the performance of the monolingual model. Requires translation pairs.

practical applications. Suppose we are in a scenario where neither strong English models nor translation pairs are available, a pretraining-based model like MLMM is preferred since we might collect image-text pairs from the Web. In scenarios where we need to extend the existing service to other languages, and we also want to maintain the performance of the existing language, a generalization-based method such as MLA is more suitable because we can leverage the strong monolingual model.

3.3 Multilingual CLIP4VLA

In this section, we propose CLIP4VLA, which extends our model from static images to dynamic videos. Since audio modality contains rich semantic information of the video[1, 40], we firstly extend CLIP[4] to support audio modality by introducing an audio encoder. After that, we extend the CLIP4VLA model to multilingual by adopting the generalization-based methods described in Section 3.2. Finally, we evaluate the models on multilingual video-text retrieval benchmark.

3.3.1 CLIP4VLA

To maintain consistency with the text encoder and visual encoder in model structure and parameter weights, we adopt an audio encoder with the same architecture as the vision encoder (i.e., ViT-B/32 of CLIP). To process the audio signal into image format, we first transfer the audio into a continuous spectrogram of $t \times H$, and then cut it into images of $k \times W \times H$ along the time dimension without overlapping.

Due to the temporal nature of audio, we use video corpus instead of images to pretrain the audio encoder. The first step is to encode the three modalities using the three encoders. Concretely, given a batch of videos V with corresponding annotations T and audios A, we denote the text sequence of T_i as $\{t_i^1, t_i^2, \cdots, t_i^{L_T}\}$, the frame sequence of V_i as $\{v_i^1, v_i^2, \cdots, v_i^{L_V}\}$, and the audio spectragram sequence of A_i as $\{a_i^1, a_i^2, \cdots, a_i^{L_A}\}$, where L refers to the sequence length of each modality. Afterwards, we follow CLIP[4] to encode T_i into word embedding sequence $X_i = \{x_i^1, x_i^2, \cdots, x_i^{L_T}\}$, and take the output of [EOS] token as the global sentence embedding x_i^g. Similarly, the frame sequence V_i is encoded into $Y_i = \{y_i^1, y_i^2, \cdots, y_i^{L_V}\}$ and the audio spectragram sequence A_i is encoded into $Z_i = \{z_i^1, z_i^2, \cdots, z_i^{L_A}\}$. By average pooling Y_i and Z_i, we obtain the global visual embedding y_i^g and audio embedding z_i^g, respectively.

To pretrain the additional audio encoder, we propose two pretraining objectives: inter-modal contrastive learning and intra-modal contrastive learning. The former objective aims to learn the cross-modal alignment between audio and the other modalities, and the latter aims to learn the inherent characteristics of audios themselves. For inter-modal contrastive learning, we conduct the audio-text similarity matrix in $B \times B$ by computing the cosine similarities of all audio-text pairs within a batch of data with batch size B. The audio encoder is trained to align audio with text by maximizing the similarity scores of B corresponding pairs while minimizing the other $B \times (B-1)$ pairs. Similarly, the audio-vision similarity matrix is computed for the audio encoder to align audio with the vision modality. For intra-modal contrastive learning, we follow SimCLR[41] to align original audio with augmented ones. Specifically, we randomly mask the original audio spectrogram a_i to $\hat{a}_i$ along both the channel and temporal dimensions, and obtain the global embedding of augmented audio $\hat{a}_i^g$ through audio encoder followed by mean pooling. Similar to inter-modal contrastive learning, the augmented audio embedding $\hat{a}_i^g$ is treated as the positive sample of a_i^g and the other masked audios within a batch are negative samples.

Since only the audio encoder is optimized during training to align audio modality with text and vision modalities, the text encoder and image encoder retain the original cross-modal knowledge of CLIP. Thus, we could further empower CLIP4VLA with multilingual capability similar to multilingual CLIP (Section 3.2.1) or multilingual acquisition (Section 3.2.2), namely multilingual CLIP4VLA (MCLIP4VLA) and ACLIP4VLA, respectively.

3.3.2 Pretraining datasets

We pretrain our proposed CLIP4VLA on large scale datasets including Howto100M[42] and Audioset[43]. Howto100M contains over 1 million tutorial videos crawling from YouTube, and is post-processed into 120 million text-video clips. Audioset contains 2 million video clips in total, and each clip is manually labeled with its event classes such as "raining, backing". To generate coherent sentences paired with the video in Audioset, we use the

template "The sound of __, __, ..." and fill in the blanks with event annotations. After filtering out the videos with missing audio and unmatched audio, we finally get 100 million Howto100M text-video pairs with narrated audio and 1.6 million text-video pairs with ambient audio. To address the unbalanced data quantity of these two datasets, we keep the 1:1 data ratio of these two datasets within a batch.

3.3.3 Implementation details

For pretraining, to make full use of existing text-visual knowledge, we initialize the text and vision backbone with CLIP, and initialize the audio backbone with the vision backbone. The pretraining batch size is set to 256 and the learning rate is set to 1×10^{-6}. The whole pretraining process takes 96 V100 days. Since the text and vision encoders have been well pretrained, we freeze these two encoders during the pretraining of the audio encoder. For fine-tuning, we replace the text backbone of CLIP with multilingual encoder (MBERT for MCLIP and MLA for ACLIP). For MCLIP4VLA, we randomly sample sentence scripts from different languages to reduce the training cost. The whole training procedure takes 5 468 steps (5 epochs), with a batch size of 128 and a learning rate of 1×10^{-7}. For ACLIP4VLA, we first finetune three backbones with only English scripts for 5 epochs, and then use multilingual scripts to finetune the acquirers of different languages. The acquirer fine-tuning takes another 5 epochs with a learning rate of 1×10^{-5}. We add the visual and audio features together for retrieval. Another concern is that not all videos contain audio information in MSR-VTT. To address the missing modality problem for each silent video during testing, we pair it with the audio of the most similar video, which is chosen from the corresponding training set according to the cosine similarity of global vision features.

3.3.4 Experiments

We evaluate our proposed multilingual multimodal pretrained model for video retrieval on MSR-VTT-7k[37, 44]. MSR-VTT-7k contains 7K videos for training and 1K videos for testing, and each video is paired with 20 descriptions per language. The dataset includes a total of 9 languages. English is the original language, and the other 8 languages are translated from English. The models to be compared include MCLIP, ACLIP, MCLIP4VLA, and ACLIP4VLA.

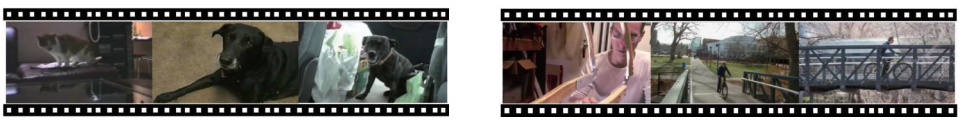
As shown in Table 8, our proposed methods outperform the existing works by a large margin. We consider that the major improvement comes from CLIP. By utilizing the well pretrained visual backbone, the generalization-based methods can boost the performance on all languages. Specifically, we observe that introducing audio modality makes further improvement by approximately 1 point of Recall@1 (comparing row 4 to row 2, row 5 to row 3). We consider that the improvement comes from the audio providing complementary information to the vision modality. For example, some text queries contain descriptions of natural sounds (Fig. 4(a)) and verbal contents (Fig. 4(b)) in the videos. It can be difficult to capture these two types of information solely based on the vision modality. Since the pretraining dataset of CLIP4VLA contains both speech (HowTo100M) and ambient sound (AudioSet), it can leverage the audio modality to improve the performance of audio-related queries.

4 Conclusions

In this paper, we explore two methods to achieve multimodal and multilingual capability: the pretraining-based methods and the generalization-based methods. For the pretraining-based methods, we propose MLMM that is pretrained on large-scale multilingual image-text datasets. For the generalization-based methods, we adopt MKD and MLA to the monolingual pretraining model CLIP. Experimental results show that MLMM achieves state-of-the-art performance on multilingual image-text retrieval, multilingual VQA and multilingual image captioning,

Table 8 Multilingual video-text retrieval results on MSR-VTT-7k

	Method	EN	ZH	CS	DE	ES	FR	RU	SW	VI	Avg.
1	SOTA[22]	23.1	20.0	21.8	21.1	21.9	21.8	20.5	14.4	10.9	19.4
2	M-CLIP	35.0	29.9	31.8	33.1	32.6	32.8	30.3	21.5	14.1	29.0
3	MLA_{CLIP}	42.5	31.6	30.5	33.1	33.5	34.5	28.9	24.3	16.9	30.6
4	MCLIP4VLA	35.8	30.4	**32.7**	**33.7**	33.5	33.8	**32.0**	22.6	15.9	30.0
5	ACLIP4VLA	**43.1**	**32.3**	31.3	33.4	**34.7**	**34.9**	30.0	**25.9**	**18.2**	**31.5**



(a) Query: A dog barks at a cat (b) Query: He says "bikes are not just a novelty item"

Fig. 4 Example queries in MSRVTT[44] search for (a) natural sounds and (b) verbal information in the video.

demonstrating convincing cross-lingual transfer ability. Both M-CLIP and MLA can achieve comparable results with much less computing cost and multilingual training data. The advantages and disadvantages of both methods are discussed as well to provide suggestions in practical applications for multilingual models.

Acknowledgements

This work was partially supported by the National Natural Science Foundation of China (No. 62072462), the National Key R&D Program of China (No. 2020AAA0 108600), and the Large-scale Pretraining Program 468 of Beijing Academy of Artificial Intelligence (BAAI).

References

[1] H. Zhu, M. D. Luo, R. Wang, A. H. Zheng, R. He. Deep audio-visual learning: A survey. *International Journal of Automation and Computing*, vol. 18, no. 3, pp. 351–376, 2021. DOI: 10.1007/s11633-021-1293-0.

[2] L. W. Zhou, H. Palangi, L. Zhang, H. D. Hu, J. Corso, J. F. Gao. Unified vision-language pre-training for image captioning and VQA. In *Proceedings of AAAI Conference on Artificial Intelligence*, vol. 34, no. 7, pp. 13041–13049, 2020. DOI: 10.1609/aaai.v34i07.7005.

[3] Y. C. Chen, L. J. Li, L. C. Yu, A. El Kholy, F. Ahmed, Z. Gan, Y. Cheng, J. J. Liu. UNITER: Universal image-text representation learning. In *Proceedings of the 16th European Conference on Computer Vision*, Springer, Glasgow, UK, pp. 104–120, 2020. DOI: 10.1007/978-3-030-58577-8_7.

[4] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, I. Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, 2021.

[5] N. Reimers, I. Gurevych. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of Conference on Empirical Methods in Natural Language Processing*, pp. 4512–4525, 2020. DOI: 10.18653/v1/2020.emnlp-main.365.

[6] L. Zhang, A. W. Hu, Q. Jin. Generalizing multimodal pre-training into multilingual via language acquisition. [Online], Available: https://arxiv.org/abs/2206.11091, 2022.

[7] G. Hinton, O. Vinyals, J. Dean. Distilling the knowledge in a neural network. [Online], Available: https://arxiv.org/abs/1503.02531, 2015.

[8] M. H. Ni, H. Y. Huang, L. Su, E. Cui, T. Bharti, L. J. Wang, D. D. Zhang, N. Duan. M3P: Learning universal representations via multitask multilingual multimodal pre-training. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Nashville, USA, pp. 3977–3986, 2021. DOI: 10.1109/CVPR46437.2021.00397.

[9] M. Y. Zhou, L. W. Zhou, S. H. Wang, Y. Cheng, L. J. Li, Z. Yu, J. J. Liu. UC^2: Universal cross-lingual cross-modal vision-and-language pre-training. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Nashville, USA, pp. 4155–4165, 2021. DOI: 10.1109/CVPR46437.2021.00414.

[10] A. Jain, M. Guo, K. Srinivasan, T. Chen, S. Kudugunta, C. Jia, Y. F. Yang, J. Baldridge. MURAL: Multimodal, multitask representations across languages. In *Proceedings of Findings of the Association for Computational Linguistics*, Punta Cana, Dominican Republic, pp. 3449–3463, 2021. DOI: 10.18653/v1/2021.findings-emnlp.293.

[11] S. K. Vipparthi, S. K. Nagar. Local extreme complete trio pattern for multimedia image retrieval system. *International Journal of Automation and Computing*, vol. 13, no. 5, pp. 457–467, 2016. DOI: 10.1007/s11633-016-0978-2.

[12] Y. Q. Huo, M. L. Zhang, G. Z. Liu, H. Y. Lu, Y. Z. Gao, G. X. Yang, J. Y. Wen, H. Zhang, B. G. Xu, W. H. Zheng, Z. Z. Xi, Y. Q. Yang, A. W. Hu, J. M. Zhao, R C. Li, Y. D. Zhao, L. Zhang, Y. Q. Song, X. Hong, W. Q. Cui, D. Y. Hou, Y. Y. Li, J. Y. Li, P. Y. Liu, Z. Gong, C. H. Jin, Y. C. Sun, S. Z. Chen, Z. W. Lu, Z. C. Dou, Q. Jin, Y. Y. Lan, W. X. Zhao, R. H. Song, J. R. Wen. WenLan: Bridging vision and language by large-scale multi-modal pre-training. [Online], Available: https://arxiv.org/abs/2103.06561, 2021.

[13] P. Sharma, N. Ding, S. Goodman, R. 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*, Melbourne, Australia, pp. 2556–2565, 2018. DOI: 10.18653/v1/P18-1238.

[14] S. Changpinyo, P. Sharma, N. Ding, R. Soricut. Conceptual 12M: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Nashville, USA, pp. 3558–3568, 2021. DOI: 10.1109/CVPR46437.2021.00356.

[15] N. Y. Fei, Z. W. Lu, Y. Z. Gao, G. X. Yang, Y. Huo, J.Y. Wen, H.Y. Lu, R. H. Song, X. Gao, T. Xiang, H. Sun, J-R. Wen. WenLan 2.0: Make AI imagine via a multimodal foundation model. [Online], Available: http://hdl.handle.net/10754/673094, 2021.

[16] N. Houlsby, A. Giurgiu, S. Jastrzebski, B. Morrone, Q. De Laroussilhe, A. Gesmundo, M. Attariyan, S. Gelly. Parameter-efficient transfer learning for NLP. In *Proceedings of the 36th International Conference on Machine Learning*, Long Beach, USA, pp. 2790–2799, 2019.

[17] J. Pfeiffer, I. Vulić, I. Gurevych, S. Ruder. MAD-X: An adapter-based framework for multi-task cross-lingual transfer. In *Proceedings of Conference on Empirical Methods in Natural Language Processing*, pp. 7654–7673, 2020. DOI: 10.18653/v1/2020.emnlp-main.617.

[18] P. Anderson, X. D. He, C. Buehler, D. Teney, M. Johnson, S. Gould, L. Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Salt Lake City, USA, pp. 6077–6086, 2018. DOI: 10.1109/CVPR.2018.00636.

[19] X. L. Zou, T. J. Huang, S. Wu. Towards a new paradigm for brain-inspired computer vision. *Machine Intelligence Research*, vol. 19, no. 5, pp. 412–424, 2022. DOI: 10.1007/s11633-022-1370-z.

[20] T. Kudo, J. Richardson. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, Association for Computational Linguistics, Brussels, Belgium, pp. 66–71, 2018. DOI: 10.18653/v1/D18-2012.

[21] J. Devlin, M. W. Chang, K. Lee, K. 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*, Minneapolis, Minnesota, pp. 4171–4186, 2019. DOI: 10.18653/v1/N19-1423.

[22] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, pp. 8440–8451, 2020. DOI: 10.18653/v1/2020.acl-main.747.

[23] R. Krishna, Y. K. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L. J. Li, D. A. Shamma, M. S. Bernstein, F. F. Li. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, vol. 123, no. 1, pp. 32–73, 2017. DOI: 10.1007/s11263-016-0981-7.

[24] K. M. He, X. Y. Zhang, S. Q. Ren, J. Sun. Deep residual learning for image recognition. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, Las Vegas, USA, pp. 770–778, 2016. DOI: 10.1109/CVPR.2016.90.

[25] D. Elliott, S. Frank, K. Sima′an, L. Specia. Multi30k: Multilingual English-German image descriptions. In *Proceedings of the 5th Workshop on Vision and Language*, Berlin, Germany, pp. 70–74, 2016. DOI: 10.18653/v1/W16-3210.

[26] X. L. Chen, H. Fang, T. Y. Lin, R. Vedantam, S. Gupta, P. Dollar, C. L. Zitnick Microsoft COCO captions: Data collection and evaluation server. [Online], Available: https://arxiv.org/abs/1504.00325, 2015.

[27] A. Karpathy, L. Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, Boston, USA, pp. 3128–3137, 2015. DOI: 10.1109/CVPR.2015.7298932.

[28] Y. Yoshikawa, Y. Shigeto, A. Takeuchi. STAIR captions: Constructing a large-scale Japanese image caption dataset. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, Vancouver, Canada, pp. 417–421, 2017. DOI: 10.18653/v1/P17-2066.

[29] X. R. Li, C. X. Xu, X. X. Wang, W. Y. Lan, Z. X. Jia, G. Yang, J. P. Xu. COCO-CN for cross-lingual image tagging, captioning, and retrieval. *IEEE Transactions on Multimedia*, vol. 21, no. 9, pp. 2347–2360, 2019. DOI: 10.1109/TMM.2019.2896494.

[30] Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, D. Parikh. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, USA, pp. 6325–6334, 2017. DOI: 10.1109/CVPR.2017.670.

[31] N. Shimizu, N. Rong, T. Miyazaki. Visual question answering dataset for bilingual image understanding: A study of cross-lingual transfer using attention maps. In *Proceedings of the 27th International Conference on Computational Linguistics*, Santa Fe, USA, pp. 1918–1928, 2018.

[32] M. Johnson, M. Schuster, Q. V. Le, M. Krikun, Y. H. Wu, Z. F. Chen, N. Thorat, F. Viégas, M. Wattenberg, G. Corrado, M. Hughes, J. Dean. Google′s multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, vol. 5, pp. 339–351, 2017. DOI: 10.1162/tacl_a_00065.

[33] S. Tsutsui, D. Crandall. Using artificial tokens to control languages for multilingual image caption generation. [Online], Available: https://arxiv.org/abs/1706.06275, 2017.

[34] K. Papineni, S. Roukos, T. Ward, W. J. Zhu. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACM, Philadelphia Pennsylvania, pp. 311–318, 2002. DOI: 10.3115/1073083.1073135.

[35] R. Vedantam, C. L.Zitnick, D. Parikh. CIDEr: Consensus-based image description evaluation. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Boston, USA, pp. 4566–4575, 2015. DOI: 10.1109/CVPR.2015.7299087.

[36] F. Mitzalis, O. Caglayan, P. Madhyastha, L. Specia. BERTGen: Multi-task generation through BERT. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pp. 6440–6455, 2021. DOI: 10.18653/v1/2021.acl-long.503.

[37] P. Y. Huang, M. Patrick, J. J. Hu, G. Neubig, F. Metze, A. Hauptmann. Multilingual multimodal pre-training for zero-shot cross-lingual transfer of vision-language models. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 2443–2459, 2021. DOI: 10.18653/v1/2021.naacl-main.195.

[38] F. Carlsson, P. Eisen, F. Rekathati, M. Sahlgren. Cross-lingual and multilingual CLIP. In *Proceedings of the 13th Language Resources and Evaluation Conference*, Marseille, France, pp. 6848–6854, 2022.

[39] F. Faghri, D. J. Fleet, J. R. Kiros, S. Fidler. VSE$_{++}$: Improving visual-semantic embeddings with hard negatives. In *Proceedings of British Machine Vision Conference*, Newcastle, UK, 2018. [Online], Available: http://bmvc2018.org/contents/papers/0344.pdf.

[40] L. F. Wu, Q. Wang, M. Jian, Y. Qiao, B. X. Zhao. A comprehensive review of group activity recognition in videos. *International Journal of Automation and Computing*, vol. 18, no. 3, pp. 334–350, 2021. DOI: 10.1007/s11633-020-1258-8.

[41] T. Chen, S. Kornblith, M. Norouzi, G. E. Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 1597–1607, 2020.

[42] A. Miech, D. Zhukov, J. B. Alayrac, M. Tapaswi, I. Laptev, J. Sivic. HowTo100M: Learning a text-video embedding by watching hundred million narrated video clips. In *Proceedings of IEEE/CVF International Conference on Computer Vision*, IEEE, Seoul, Republic of Korea, pp. 2630–2640, 2019. DOI: 10.1109/ICCV.2019.00272.

[43] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, M. Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, New Orleans, USA, pp. 776–780, 2017. DOI: 10.1109/ICASSP.2017.7952261.

[44] J. Xu, T. Mei, T. Yao, Y. Rui. MSR-VTT: A large video description dataset for bridging video and language. In *Proceedings of IEEE Conference on Computer Vision and*

Pattern Recognition, Las Vegas, USA, pp. 5288–5296, 2016. DOI: 10.1109/CVPR.2016.571.

Liang Zhang received the B. Sc. degree in computer science and technology from China University of Mining and Technology, China in 2020. He is currently a Ph. D. degree candidate in big data science and engineering at School of Information, Renmin University of China, China.

His research interests include multilingual machine learning, cross-modal retrieval and multimodal reading comprehension.

E-mail: zhangliang00@ruc.edu.cn

ORCID iD: 0000-0002-6187-3628

Ludan Ruan received the B. Sc. degree in computer science and technology from Renmin University of China, China in 2020. She is currently a master student in computer application technology at School of Information, Renmin University of China, China.

Her research interests include multimodal pretraining, temporal sentence grounding and video generation.

E-mail: ruanld@ruc.edu.cn

Anwen Hu received the B. Sc. degree in computer science and technology from Renmin University of China, China in 2017. He is currently a Ph. D. degree candidate in big data science and engineering at School of Information, Renmin University of China, China.

His research interests include image captioning, vision and language, natural language processing.

E-mail: anwenhu@ruc.edu.cn

Qin Jin received the Ph. D. degree in language and information technologies from Carnegie Mellon University, USA in 2007. She is a full professor at School of Information, Renmin University of China, China. She has published in various top-tier conferences related to computer vision (CV), multimedia, and AI, served as technical program chair, area chair, program committee member, and reviewer and associate editor for conferences and journals.

Her research interests include multimedia content analysis, human computer interaction and machine learning in general.

E-mail: qjin@ruc.edu.cn (Corresponding author)

ORCID iD: 0000-0001-6486-6020