Cross-lingual Cross-modal Pretraining for Multimodal Retrieval

Hongliang Fei, Tan Yu, Ping Li
Cognitive Computing Lab
Baidu Research
10900 NE 8th St, Bellevue, WA 98004, USA
{hongliangfei, tanyu01, liping11}@baidu.com

Abstract

Recent pretrained vision-language models have achieved impressive performance on cross-modal retrieval tasks in English. Their success, however, heavily depends on the availability of many annotated image-caption datasets for pretraining, where the texts are not necessarily in English. Although we can utilize machine translation (MT) tools to translate non-English text to English, the performance still largely relies on MT's quality and may suffer from high latency problems in real-world applications. This paper proposes a new approach to learn cross-lingual cross-modal representations for matching images and their relevant captions in multiple languages. We seamlessly combine cross-lingual pretraining objectives and cross-modal pretraining objectives in a unified framework to learn image and text in a joint embedding space from available English image-caption data, monolingual and parallel corpus. We show that our approach achieves SOTA performance in retrieval tasks on two multimodal multilingual image caption benchmarks: Multi30k with German captions and MSCOCO with Japanese captions.

1 Introduction

Recent pretrained vision-language models (Chen et al., 2020; Li et al., 2020; Su et al., 2020; Gan et al., 2020; Luo et al., 2020) based on Transformer (Vaswani et al., 2017) have achieved remarkable performance on cross-modal retrieval (Li et al., 2020; Yu et al., 2020, 2021b), image captioning (Chen et al., 2020) and visual question and answering (VQA) (Su et al., 2020) tasks in English. For instance, most leading competitors in the VQA contest[1] rely on the transformer-based pretrained vision-language models.

However, their success heavily depends on the availability of a large amount of annotated image-caption pretraining datasets (e.g., conceptual captions (Sharma et al., 2018)). In reality, there are limited such data in other languages. When generalizing to cross-lingual cross-modal downstream tasks, a straightforward way is to utilize machine translation (MT) tools to translate non-English text to English and reuse the fine-tuned models in English. Nevertheless, the performance strongly relies on the MT tool's capability and suffers from high latency problems in real-world applications.

To learn multilingual multimodal representations, recent researchers utilized multilingual datasets to model images and text captions in a joint embedding space. Based on how the shared feature space is learned, there are two categories: word-level alignments (Mohammadshahi et al., 2019) and sentence-level alignments (Wehrmann et al., 2019; Rajendran et al., 2016). Those models can capture a certain level of semantic similarity among languages and images. They, however, only modeled the relevance of text and images in a global manner. Such a limitation may prevent these models from effectively detecting relevance locally.

In parallel, cross-lingual language models such as multilingual BERT (Devlin et al., 2019) and XLM (Conneau and Lample, 2019), and pretrained vision-language models (Chen et al., 2020; Li et al., 2020; Su et al., 2020) have been prevalent in bridging different languages and modalities. Those models use the Transformer (Vaswani et al., 2017) architecture simultaneously pretrained from multiple languages or image-caption pairs to construct an encoder, and then fine-tune the encoder on downstream applications with task-specific objectives. The whole process enables sufficient interaction across languages and other modalities via cross-attention. However, current cross-lingual models and cross-modal models are trained separately on multilingual corpus and English-caption data. Hence the resulting pretrained models are not directly applicable to downstream cross-modal tasks involving non-English languages.

[1] https://visualqa.org/roe.html

Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 3644–3650
June 6–11, 2021.

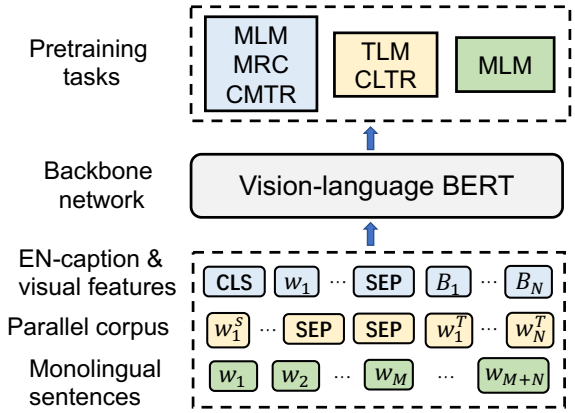

Figure 1: Illustration of the proposed pretraining model. Our input data consists of three sources: English captions and corresponding visual bounding box features, parallel sentences involving English and other languages, and monolingual text corpus. Each data source is associated with one or more pretraining tasks as indicated by the same color. The acronyms for pretraining tasks are summarized in Table 1.

MLM	Masked language modeling task
TLM	Translation language modeling task
MRC	Masked region classification task
CLTR	Cross-lingual text recovery task
CMTR	Cross-modal text recovery task in this paper

Table 1: Frequently used acronyms in this paper.

This paper proposes a cross-lingual cross-modal pretraining framework to learn a language invariant representation across image and text modalities. We hypothesize that introducing pretraining tasks involving different languages and modalities and modeling the interaction among them leads to a more powerful joint representation and generalizes well to downstream tasks. Extending previous vision-language pretraining works (e.g., Su et al. (2020)) that learn parameters solely based on the English-image caption data, we introduce monolingual and parallel corpus involving other languages to refine the shared latent space further.

In Figure 1, we provide a skeleton of our pretraining framework, which is built on top of vision-language BERT models (Su et al., 2020; Li et al., 2020) with more pretraining tasks and data sources. In particular, we use masked language modeling (MLM) (Devlin et al., 2019) on monolingual text corpus, and translation language modeling (TLM) adopted from XLM (Conneau and Lample, 2019) on parallel text corpus. We follow the standard vision-language pretraining models for the English-image data and use MLM on text captions and masked region classification (MRC) on image regions. Besides, motivated by the success of the cross-lingual text recovery (CLTR) task in Unicoder (Huang et al., 2019), we propose a cross-modal text recovery (CMTR) task. Like CLTR, CMTR leverages the attention matrix between image-caption pairs to learn the alignment among words and regions of interest in images.

We performed text-to-image and image-to-text retrieval tasks on two multimodal multilingual image caption benchmarks: Multi30k (German and English) captions and MSCOCO (English and Japanese). We achieve SOTA results on retrieval tasks involving Japanese and German languages, compared with a machine translation baseline and other recently published works.

2 Related Work

2.1 Vision-language Pretrained Model

Recently, BERT (Devlin et al., 2019) based vision-language pretraining models (Chen et al., 2020; Li et al., 2020; Su et al., 2020; Gan et al., 2020; Luo et al., 2020) emerge. In those models, the pretraining typically consists of three types of tasks: 1) masked language modeling, 2) masked region modeling, and 3) text-image matching. By exploiting the cross-modal attention and being pretrained on large-scale datasets, cross-modal BERT methods have achieved state-of-the-art performance in many text-vision understanding tasks. Nevertheless, all the above models deal with a single language English and image or video domain.

2.2 Cross-lingual Pretrained Model

Cross-lingual pretrained language models (Devlin et al., 2019; Conneau and Lample, 2019) are capable of simultaneously encoding texts from multiple languages. Most notably, multilingual BERT (Devlin et al., 2019) takes the same model structure and training objective as BERT but was pretrained on more than 100 languages on Wikipedia. XLM model (Conneau and Lample, 2019) is pretrained with MLM and TLM to take advantage of parallel sentence resources if available. Evaluations on a series of cross-lingual transfer tasks (Fei and Li, 2020; Yu et al., 2021a) have shown that these cross-lingual LMs have significant utilities for transferring knowledge between languages. Therefore, we propose integrating cross-lingual pretraining tasks with vision-language pretraining to obtain a universal multilingual multimodal representation.

3 Methodology

Our framework adopts the network structure of VL-BERT (Su et al., 2020). VL-BERT is a single-stream cross-modal model that concatenates word features from the text and bounding box features from the image and feeds the concatenated sequence into a series of transformer blocks.

3.1 Pretraining tasks

Both vision-grounded masked language model (MLM) and text-grounded masked region classification (MRC) task on image-caption data are used in our model by default, as they have shown strong performance in VL-BERT (Su et al., 2020; Li et al., 2020). Since we introduce auxiliary multilingual text corpus, we also use MLM on the texts in other languages by default. Motivated by Unicoder (Huang et al., 2019) showing that pretrained models can be further improved by involving more tasks, we introduce two additional cross-lingual pretraining tasks and one cross-modal task for improving the performance.

Cross-model Text Recovery. This task (CMTR) is motivated by the multilingual pretraining model Unicoder (Huang et al., 2019). As shown in Figure 2, CMTR is based on the image-caption pairs as input, but it does not use the original caption words. Instead, it computes an alignment between word features and bounding box features extracted by tools (e.g., Faster-RCNN (Anderson et al., 2018)), and uses attended features to simultaneously recover all input words. In particular, let $(\mathbf{B}, \mathbf{E})$ be an image-caption input pair, where $\mathbf{B} = (\mathbf{b}_1, \mathbf{b}_2, \cdots, \mathbf{b}_n)$ are bounding box feature embeddings and $\mathbf{E} = (\mathbf{e}_1, \mathbf{e}_2, \cdots, \mathbf{e}_m)$ are word embeddings. CMTR first calculates an attended representation for the caption words with bounding box features as $\hat{\mathbf{e}}_i = \sum_{j=1}^{n} \tilde{a}_{ij}\mathbf{b}_j$, where $\tilde{a}_{ij} = \mathrm{softmax}(A_{i,:})[j]$, $\mathbf{b}_j \in \mathcal{R}^h$, $\mathbf{e}_i \in \mathcal{R}^h$, and h denotes the embedding dimension. $\mathbf{A} \in \mathcal{R}^{m \times n}$ is the attention matrix calculated by bi-linear attention as $A_{ij} = \mathbf{e}_i^T \mathbf{W} \mathbf{b}_j$, where $\mathbf{W}$ is a trainable parameter. Finally we take $\hat{\mathbf{E}} = \tanh((\hat{\mathbf{e}}_1, \hat{\mathbf{e}}_2, \cdots, \hat{\mathbf{e}}_m))$ as input and predict the original caption words. The objective function is:

$$l(X; e, d) = \mathbb{E}_{x \sim X}[\Delta(x, d(e(x)))] \quad (1)$$

where $\Delta(.,.)$ is the sum of token-level cross-entropy loss and $e(.)$ is the encoder component including the input layer, the attention layer and transformer layers. $d(.)$ is the decoder applied on the output of transformers, which is a shared linear projection layer with other MLM tasks and CLTR task introduced below.

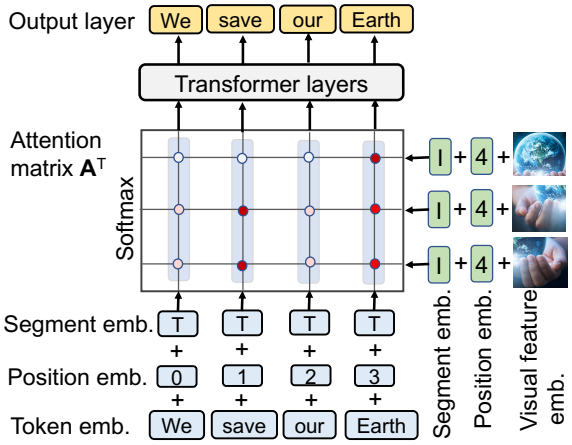

Figure 2: Cross-modal text recovery. CMTR directly learns the underlying alignments between words and regions of interest in images and generates an attended input to stacked transformer layers to recover all input words. Note that the attention matrix is transposed.

Cross-lingual Text Recovery. This task (CLTR) is adopted from Unicoder (Huang et al., 2019), which takes a pair of parallel sentences (X, Y) and lets the pretrained model learn the underlying word alignments between two languages. Similar to CMTR, we also use the bi-linear attention mechanism to compute an attended representation $\hat{X}$ for input sentence X in the source language with its parallel sentence Y, and then try to recover X using the attended input $\hat{X}$. In CLTR task, we optimize the same objective function in Eq. (1). Note that CLTR and CMTR do not share attention parameters since there is still a large modal gap between text and image before applying cross-attention.

Translation Language Model. This task (TLM) is adopted from XLM (Conneau and Lample, 2019), which takes a pair of parallel sentences with randomly masked tokens in different languages as input. The model is trained to predict the masked tokens by attending to local contexts and distant contexts in another language. Interested readers please refer to Conneau and Lample (2019) for more details about its objective function.

3.2 Fine-tuning for Cross-modal Retrieval

For fine-tuning, we minimize the triplet ranking loss to fine-tune the retrieval model. To boost the performance, we use the hard negative mining strategy in SCAN (Lee et al., 2018). For each text query,

there is only one positive image sample and the rest are negative. Denoting a mini-batch of training samples by $\{(q_i, I_i)\}_{i=1}^{K}$, where a query q_i is only relevant with the image I_i, we only penalize the hardest negative image in the mini-batch by

$$\mathcal{L}(q_i) = \max_{j \neq i}[R(q_i, I_j) - R(q_i, I_i) + m]_+,$$

where m is the margin set to 0.2 by default, and $[x]_+ = \max(0, x)$ is a clip function. $R(q, I)$ is the function to evaluate the similarity between query q and image I parameterized by $\mathbf{u}$ and b:

$$R(q, I) = \mathbf{u}^\top \text{BERT}_{\text{CLS}}(q, I) + b.$$

On the other hand, for each image, we only penalize the hardest negative query in the mini-batch:

$$\mathcal{L}(I_i) = \max_{j \neq i}[R(q_j, I_i) - R(q_i, I_i) + m]_+.$$

Considering the whole mini-batch of images and texts, the final loss function is computed by $\mathcal{L} = \frac{1}{K}\sum_{i=1}^{K}[\mathcal{L}(q_i) + \mathcal{L}(I_i)]$.

4 Experiment

For pretraining, we utilize two public English image-caption datasets: SBU Captions (Ordonez et al., 2011) and Conceptual Captions (Sharma et al., 2018). Due to broken URLs, we only collected around 3.7M text-image pairs in total. For monolingual (en, de, ja) text and parallel corpus (en-de), we randomly sample 20M sentences from Wikipedia text[2] and 9M parallel sentences from MultiUN corpus[3]. We also collected 2.8M en-ja parallel sentences from Pryzant et al. (2018).

For fine-tuning, we use two multilingual multimodal benchmarks for retrieval, MSCOCO (en, ja) (Lin et al., 2014) and Multi30k (en, de) (Elliott et al., 2016). MSCOCO contains 123,287 images, and each image contains five captions. Following the settings in Faghri et al. (2018), we split the English data into 113,287 training samples, 5,000 validation samples, and 5,000 testing samples. Miyazaki and Shimizu (2016) generated the Japanese captions for a subset of 33,745 images. Similarly, we split 23,745 samples for training, 5,000 for validation as 5,000 for testing. Multi30K contains 31,783 images, with each having five captions as well. Following Karpathy and Li (2015), we split the dataset into 29,783 training samples, 1,000 validation samples and 1,000 testing samples. We use R@K (K = 1,5,10) as evaluation metrics. R@K is the percentage of ground-truth matchings appearing in the top K-ranked results.

4.1 Experiment Setting

We use the multilingual BERT uncased version (Devlin et al., 2019) to initialize our model, which has 12 layers of Transformer blocks. Each block has 768 hidden units, 12 self-attention heads, and the vocabulary size is 105,879. The maximum sequence length is set to 64. Following Li et al. (2020), we detect 100 bounding boxes per image using Faster-RCNN (Anderson et al., 2018) pretrained on Visual Genome (Krishna et al., 2017).

Our pretraining is conducted on 16 NVIDIA V100 GPUs (16GB memory), and fine-tuning is conducted on 8 NVIDIA V100 GPUs. We use FP16 to speed up training and reduce memory usage. We use Adam optimizer (Kingma and Ba, 2015) and set the batch size per GPU to 16. The initial learning rate is 1e-5. We pretrain the model for 50 epochs and fine-tune the retrieval model based on the average of R@{1,5,10} on the validation set. We repeat our experiments five times and report the average metrics on the test set.

4.2 Baselines

We compare our models with several recent competitive methods. VL-BERT (Su et al., 2020) and Unicoder-VL (Li et al., 2020) are two well-known vision-language BERT based models. For VL-BERT, We reproduce the English results by fine-tuning their official pretrained model[4] and generate non-English results from their released code following the same configuration as ours. For Unicoder-VL, we adopt their reported English results in the paper. Besides pretraining based models, we also compare several methods, including cross-attention based model SCAN (Lee et al., 2018), multilingual word embedding alignment-based model AME (Mohammadshahi et al., 2019) and multilingual sentence alignment-based model LIME (Wehrmann et al., 2019). We directly use SCAN, AME, and LIME's reported performance from their papers. Finally, we compare with a machine translation baseline: "Translate-test", which translates the test data in Japanese or German to English using Google Translate, and then evaluates on fine-tuned VL-BERT retrieval model in English.

[2] http://dumps.wikimedia.org/
[3] https://bit.ly/2OvI2ZD
[4] https://bit.ly/3cZTzJW

Method	MSCOCO (en)						Multi30K (en)					
	img2txt Recall@			txt2img Recall@			img2txt Recall@			txt2img Recall@		
	1	5	10	1	5	10	1	5	10	1	5	10
SCAN	72.7	94.8	98.4	58.8	88.4	94.8	67.4	90.3	95.8	48.6	77.7	85.2
Unicoder-VL	**84.3**	**97.3**	99.3	**69.7**	**93.5**	**97.2**	**86.2**	**96.3**	**99.0**	**71.5**	**90.9**	**94.9**
VL-BERT	76.4	96.8	99.2	64.1	90.9	96.3	79.8	94.9	96.8	61.8	86.4	92.1
Ours	80.5	97.1	**99.5**	65.1	91.7	96.5	80.6	94.9	97.9	63.3	87.6	92.4

Table 2: Cross-modal retrieval results (in percentage %) for English. Best results are marked in bold.

Method	MSCOCO (ja)						Multi30K (de)					
	img2txt Recall@			txt2img Recall@			img2txt Recall@			txt2img Recall@		
	1	5	10	1	5	10	1	5	10	1	5	10
SCAN	56.5	85.7	93.0	42.5	73.6	83.4	51.8	82.0	91.0	35.7	60.9	71.0
AME	55.5	87.9	95.2	44.9	80.7	89.3	40.5	74.3	83.4	31.0	60.5	70.6
LIWE	56.9	86.1	94.1	45.1	78.0	88.2	59.9	87.5	93.7	42.3	71.1	79.8
Translate-test	66.2	88.8	94.8	52.1	82.5	90.6	69.8	90.2	94.8	51.2	77.9	86.6
VL-BERT	60.3	85.9	94.5	48.4	81.7	90.5	65.7	88.0	94.0	47.4	77.0	85.4
Ours	**67.4**	**90.6**	**96.2**	**54.4**	**84.4**	**92.2**	**71.1**	**91.2**	**95.7**	**53.7**	**80.5**	**87.6**

Table 3: Cross-modal retrieval results for Japanese (MSCOCO) and German (Multi30K). Best results with statistical significance are marked in bold (one-sample t-test with $p < 0.05$).

4.3 Experimental Results

Table 2 presents the results for English tasks. Compared with Unicoder-VL (Li et al., 2020), our model performs slightly worse but obtains better results than VL-BERT. A possible reason is that Unicoder-VL is initialized with English BERT, which is specifically optimized for English.

The benefit of our model is demonstrated in Table 3 for cross-modal retrieval tasks involving non-English languages. We first observe that the machine translation baseline "Translate-test" achieves better results than VL-BERT pretrained with MLM objective only on multilingual corpus and fine-tuned in the target language, proving the importance of aligning different languages.

Moreover, the average recall of the "Translate-test" is around 1-2% lower than our method. Such results indicate that pretraining with additional cross-lingual objectives is more effective than translating the target language into English for these two benchmarks. Though combining more powerful machine translation tools and better fine-tuned English retrieval models may lead to slightly better performance, our method learns a universal representation without dependency on external machine translation tools for particular language pairs, which is more suitable for real-world applications. Finally, compared with VL-BERT (Su et al., 2020) that is only pretrained with MLM task on multilingual corpus, our additional cross-lingual pretraining tasks bring performance improvement.

	MO (en)	MO (ja)	MK (en)	MK (de)
Full Model	72.8	**60.9**	72.0	**62.4**
w/o TLM	72.6	58.9	71.9	60.9
w/o CLTR	72.8	59.3	71.9	61.1
w/o CMTR	71.2	60.2	71.1	61.5

Table 4: Ablation study on the average of R@1. Best results with statistical significance are marked in bold. MO: MSCOCO, MK: Multi30K.

4.4 Ablation Study

To understand the effect of different components, we conduct an ablation study on the test set and report the average Recall@1 in Table 4. Although cross-lingual pretraining tasks (TLM and CLTR) do not help English-related retrieval tasks much, they contribute more than 1% improvement for Japanese and German. The result is under our expectation since those tasks effectively link non-English languages with the vision domain using English as the bridge. Among all the components, CMTR consistently contributes around 1 point improvement.

5 Conclusion

In this work, we introduce multilingual corpus and three pretraining objectives to improve transformer based vision-language models for retrieval tasks. Extensive experiments demonstrate the effectiveness of our contributions on cross-modal retrieval tasks. Detailed ablation studies justify our modeling choices. Our future work is to explore the zero-shot transferring capability of our framework.

References

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6077–6086, Salt Lake City, UT.

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 *Proceedings of the 16th European Conference on Computer Vision (ECCV), Part XXX*, pages 104–120, Glasgow, UK.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 7059–7069, Vancouver, Canada.

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 (NAACL-HLT)*, pages 4171–4186, Minneapolis, MN.

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, hosted by the 54th Annual Meeting of the Association for Computational Linguistics (VL@ACL)*, Berlin, Germany.

Fartash Faghri, David J. Fleet, Jamie Ryan Kiros, and Sanja Fidler. 2018. VSE++: improving visual-semantic embeddings with hard negatives. In *Proceedings of the British Machine Vision Conference (BMVC)*, Newcastle, UK.

Hongliang Fei and Ping Li. 2020. Cross-lingual unsupervised sentiment classification with multi-view transfer learning. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 5759–5771, Online.

Zhe Gan, Yen-Chun Chen, Linjie Li, Chen Zhu, Yu Cheng, and Jingjing Liu. 2020. Large-scale adversarial training for vision-and-language representation learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, virtual.

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.

Andrej Karpathy and Fei-Fei Li. 2015. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3128–3137, Boston, MA.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*, San Diego, CA.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *Int. J. Comput. Vis.*, 123(1):32–73.

Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. 2018. Stacked cross attention for image-text matching. In *Proceedings of the 15th European Conference on Computer Vision (ECCV), Part IV*, pages 212–228, Munich, Germany.

Gen Li, Nan Duan, Yuejian Fang, Ming Gong, and Daxin Jiang. 2020. Unicoder-VL: A universal encoder for vision and language by cross-modal pre-training. In *Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI)*, pages 11336–11344, New York, NY.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: common objects in context. In *Proceedings of the 13th European Conference on Computer Vision (ECCV), Part V*, pages 740–755, Zurich, Switzerland.

Huaishao Luo, Lei Ji, Botian Shi, Haoyang Huang, Nan Duan, Tianrui Li, Xilin Chen, and Ming Zhou. 2020. Univilm: A unified video and language pre-training model for multimodal understanding and generation. *arXiv preprint arXiv:2002.06353*.

Takashi Miyazaki and Nobuyuki Shimizu. 2016. Cross-lingual image caption generation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*, page 1780–1790, Berlin, Germany.

Alireza Mohammadshahi, Rémi Lebret, and Karl Aberer. 2019. Aligning multilingual word embeddings for cross-modal retrieval task. In *Proceedings of the Beyond Vision and Language: inTEgrating Real-world kNowledge (LANTERN@EMNLP-IJCNLP)*, pages 11–17, Hong Kong, China.

Vicente Ordonez, Girish Kulkarni, and Tamara L. Berg. 2011. Im2text: Describing images using 1 million captioned photographs. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1143–1151, Granada, Spain.

Reid Pryzant, Youngjoo Chung, Dan Jurafsky, and Denny Britz. 2018. JESC: japanese-english subtitle corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC)*, Miyazaki, Japan.

Janarthanan Rajendran, Mitesh M. Khapra, Sarath Chandar, and Balaraman Ravindran. 2016. Bridge correlational neural networks for multilingual multimodal representation learning. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 171–181, San Diego, CA.

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 (ACL)*, pages 2556–2565, Melbourne, Australia.

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 *Proceedings of the 8th International Conference on Learning Representations (ICLR)*, Addis Ababa, Ethiopia.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems (NIPS)*, pages 5998–6008, Long Beach, CA.

Jonatas Wehrmann, Maurício Armani Lopes, Douglas M. Souza, and Rodrigo C. Barros. 2019. Language-agnostic visual-semantic embeddings. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 5803–5812, Seoul, Korea.

Puxuan Yu, Hongliang Fei, and Ping Li. 2021a. Cross-lingual language model pretraining for retrieval. In *Proceedings of the Web Conference (WWW)*, Ljubljana, Slovenia.

Tan Yu, Yi Yang, Yi Li, Xiaodong Chen, Mingming Sun, and Ping Li. 2020. Combo-attention network for baidu video advertising. In *Proceedings of the 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, pages 2474–2482, Virtual Event, CA, USA.

Tan Yu, Yi Yang, Yi Li, Lin Liu, Hongliang Fei, and Ping Li. 2021b. Heterogeneous attention network for effective and efficient cross-modal retrieval. In *Proceedings of the 44th International ACM SIGIR conference on research and development in Information Retrieval (SIGIR)*, Virtual Event.