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

Mingyang Zhou[1], Luowei Zhou[2], Shuohang Wang[2], Yu Cheng[2], Linjie Li[2], Zhou Yu[1], Jingjing Liu[2]
[1]University of California, Davis
[2]Microsoft Dynamics 365 AI Research
{minzhou, joyu}@ucdavis.edu
{luozhou, shuowa, yu.cheng, lindsey.li, jingjl}@microsoft.com

A. Appendix

In this supplementary materials, we present the implementation details for downstream task finetuning (Section A.1) and more ablation study on the proposed pre-training objectives (Section A.2).

A.1. Downstream Tasks Details

Multilingual Image-Text Retrieval During fine-tuning, we train and evaluate the pre-trained UC^2 on Multi30K [4, 3, 1] and MSCOCO [2, 8, 6]. When we fine-tune UC^2 on both datasets, we use batch size of 40 and sample 2 negative image-text pairs for each sampled positive image-text pair. The pre-trained model is optimized by the Adam Optimizer with the learning rate set to $1e-4$ and a linear warm-up for the first 10% of fine-tuning. For Cross-Lingual zero-shot setting, the pre-trained UC^2 is fine-tuned on English-only training data for 30K steps. For All-Language setting, we train UC^2 on all the training data in all languages for 50K steps. The finetuning is run on 8 Nvidia V100 GPUs.

Multilingual VQA When we fine-tune the pre-trained model on VQA, the output layer is formatted to output a probability distribution over a set of predefined answers. To train the pre-trained model on VQA, we apply a binary cross-entropy loss and optimize the objective with Adam optimizer. On VQA v2.0 [5], we set batch size to 10240 and the learning rate as $2e-5$, and the model is trained for 7K steps. On VQA VG Japanese [7], the model is trained for 5K steps with the batch of 5120 and the learning rate of $8e-5$. The weight decay for the fine-tuning on both datasets is set to 0.01. The fine-tuning for VQA is run on 4 Nvidia V100 GPUs.

A.2. Ablation Studies

In this section, we provide more ablation studies on our proposed pre-training objectives that could not fit in the main paper due to space limit. First, we provide evidence to show our proposed MRTM is better than the traditional masked region modeling task MRC for multilingual multimodal pre-training. Second, we study the impact of number of languages included inpre-training data. Last, we explore the effect of using different pivoting languages in our proposed VTLM task.

Pre-training tasks	ITR Meta-Ave	VQA EN	VQA JA
ITM+MLM+MRC	85.1	70.60	33.4
ITM+MLM+MRTM	**85.3**	**71.45**	**34.1**

Table 1. Direct ablation on comparison between the proposed MRTM and the MRC. The presentation of the result is simplified to only include the Meta-Average for the mutilingual image-text retrieval over both Multi30K and MSCOCO, the accuracy on VQA v2.0 test-dev split (referred as VQA EN), and the accuracy on VQA VG Japanese (referred as VQA JA).

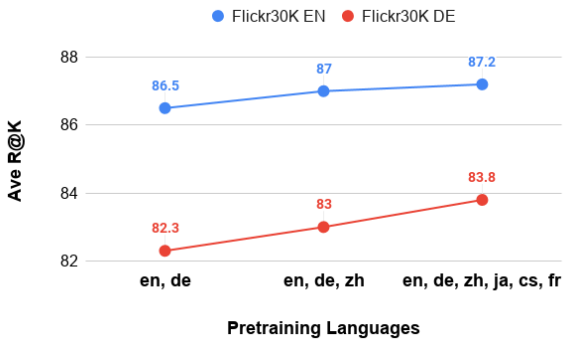

Figure 1. Comparison of image-text retrieval performance when pre-trained with different groups of languages (average R@K on Flickr30K English and German).

MRTM vs MRC For this ablation, we pre-train UC^2 with ITM, MLM and MRC and compare the results to the pre-trained UC^2 optimized with ITM, MLM and MRTM. The results is summarized in Table. 1. As shown in table 1, compare the pretrained UC^2 that employs the traditional task MRC and the one that employs our proposed MRTM,

we can see that the performance on the image-text retrieval task are similar, but MRTM leads to marginal improvement on the multilingual VQA tasks. This observation is consistent with our hypothesis that the proposed MRTM augments the local alignment between image regions and the words in different languages which benefits downstream tasks that rely on region-level recognition and reasoning.

Effect of Pre-training languages As we use machine translation models to expand the pre-training corpus, theoretically, we can have as many languages as needed. We conduct further experiments to verify the impact of number of languages included in pre-training data. We create three variants of pre-training corpus, where the number of languages are 2, 3, and 6, respectively.[1] Every corpus contains English and German. We add Chinese to construct the corpus with 3 languages, and the corpus with 6 languages contains all the languages used to pre-train our full model. The pre-trained models are evaluated on image-text retrieval task in English and German, by finetuning on target language.

Figure 1 shows that when the number of pre-training languages increases, the performance on image-text retrieval on different languages (English and German) slightly improves. This result demonstrates that cross-lingual cross-modal pre-training can effectively leverage different languages to learn stronger vision-to-monolingual-sentence alignment. Meanwhile, as we maintain the same pre-training epochs for all three experiments, we also observe that the benefit of multilingual V+L pre-training is compensating for the reduced training time allocated to each language. Although more comprehensive analysis in future study can help us better understand the trade-off between language capacity and performance on downstream tasks, our observation to some extent still suggests that our model is scalable to pre-training on a large corpus with many languages within a reasonable time frame.

Effect of Pivoting Language in VTLM We also conducted a controlled experiment to learn the effect of different pivoting languages in VTLM for the multi-lingual multi-modal pre-training. In this controlled experiment, we pre-train UC2 with all the objectives but change the pivoting language in VTLM from English to Chinese. When we evaluate the pre-trained model on the multilingual image-text retrieval task, the meta-ave score for the pre-trained model with VTLM pivoted on Chinese is dropped from 86.2 to 85.5. This to some extent suggest that English is a more optimal pivoting language to learn the cross-lingual cross-modal shared representation space. Another potential reason for the limited performance is due to the noiseness in the pre-trained Chinese captions gained via automatic machine translation. To gain more solid conclusion to determine the optimal pivoting language, more comprehensive experiments need to be conducted in the future work.

[1]For fair comparison, we constraint the training time to be the same with different pre-training corpus.

References

[1] Loïc Barrault, Fethi Bougares, Lucia Specia, Chiraag Lala, Desmond Elliott, and Stella Frank. Findings of the third shared task on multimodal machine translation. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 304–323, Belgium, Brussels, Oct. 2018. Association for Computational Linguistics. 1

[2] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO captions: Data collection and evaluation server. *CoRR*, abs/1504.00325, 2015. 1

[3] Desmond Elliott, Stella Frank, Loïc Barrault, Fethi Bougares, and Lucia Specia. Findings of the second shared task on multimodal machine translation and multilingual image description. In *Proceedings of the Second Conference on Machine Translation*, pages 215–233, Copenhagen, Denmark, Sept. 2017. Association for Computational Linguistics. 1

[4] Desmond Elliott, Stella Frank, Khalil Sima'an, and Lucia Specia. Multi30K: Multilingual English-German image descriptions. In *Proceedings of the 5th Workshop on Vision and Language*, pages 70–74, Berlin, Germany, Aug. 2016. Association for Computational Linguistics. 1

[5] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 1

[6] Xirong Li, Xiaoxu Wang, Chaoxi Xu, Weiyu Lan, Qijie Wei, Gang Yang, and Jieping Xu. COCO-CN for cross-lingual image tagging, captioning and retrieval. *CoRR*, abs/1805.08661, 2018. 1

[7] Nobuyuki Shimizu, Na Rong, and Takashi Miyazaki. Visual question answering dataset for bilingual image understanding: A study of cross-lingual transfer using attention maps. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1918–1928, Santa Fe, New Mexico, USA, Aug. 2018. Association for Computational Linguistics. 1

[8] Yuya Yoshikawa, Yutaro Shigeto, and Akikazu Takeuchi. STAIR captions: Constructing a large-scale Japanese image caption dataset. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 417–421, Vancouver, Canada, July 2017. Association for Computational Linguistics. 1