Supplementary Material of VLMO: Unified Vision-Language Pre-Training with Mixture-of-Modality-Experts

Anonymous Author(s)
Affiliation
Address
email

1 Supplementary Experiments

1.1 Ablation Study of MOME

Table 1 presents the ablation study of shared self-attention module used in MOME Transformer for encoding image patches and text tokens. We compare shared self-attention with separate self-attention, which encodes image patches and text tokens using different attention parameters on the first L−F layers. The shared self-attention used in MOME achieves better performance. The shared self-attention module helps VLMO learn the alignment of different modalities, and fuse images and text at bottom layers for classification tasks.

1.2 Global Hard Negative Mining

Different from ALBEF [1], which samples hard negatives from training examples of the single GPU (namely local hard negative mining). We perform hard negative mining from more candidates by gathering training examples of all GPUs (namely global hard negative mining). As shown in Table 2, our global hard negative mining brings significant improvements.

1.3 Evaluation on Vision Tasks

As shown in Table 3, we use VLMO as an image-only encoder and evaluate it on image classification (ImageNet [3]) and semantic segmentation (ADE20K [5]) tasks. The model also achieves competitive performance, even slightly better than the BEIT model used for the initialization of VLMO. The image resolution is 224×224 for ImageNet, and 512×512 for ADE20K. We perform intermediate fine-tuning on ImageNet-21k for all results of three models.

2 Supplementary Hyperparameters

2.1 Hyperparameters for Pre-Training

The vision-language pre-training of base-size model takes about two days using 64 Nvidia Tesla V100 32GB GPU cards, and the large-size model takes about three days using 128 Nvidia Tesla V100 32GB GPU cards.

For the text-only pre-training data, we use English Wikipedia and BookCorpus [6]. AdamW [2] optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.98$ is used to train the models. The maximum sequence length is set to 196. The batch size is 1024, and the peak learning rate is 2e-4. We set the weight decay to 0.01. For the base-size model, we train the model for 500k steps. The large-size model is trained for 200k steps.

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

Transformer	NLVR2		Flickr30k	
	dev	test-P	TR	IR
Separate Self-Attention	78.92	78.95	94.63	86.88
MoME (Shared Self-Attention)	**80.13**	**80.31**	**95.17**	**87.25**

Table 1: Ablation study of the shared self-attention module used in MOME. We experiment with separate attention on the first L−F layers, which encodes image patches and text tokens using different attention parameters.

Models	NLVR2	
	dev	test-P
Local hard negative mining [1]	77.70	77.95
Global hard negative mining	**79.54**	**79.48**

Table 2: Global hard negative mining improves the model. We perform experiments using 32 V100 GPUs. The batch size per GPU is 32, and the total batch size is 1024. Local hard negative mining samples hard negatives from training examples of the single GPU (32 examples), while global hard negative mining uses training examples gathered from all GPUs as the candidates (1024 examples).

2.2 Hyperparameters for Vision-Language Classification Fine-Tuning

Visual Question Answering (VQA) We fine-tune the models for 10 epochs with 128 batch size. The peak learning rate is 3e-5 for the base-size model, and 1.5e-5 for the large-size model. Following SimVLM [4], the input image resolution is 480×480. For VLMO-Large++, we use 768×768 image resolution.

Natural Language for Visual Reasoning (NLVR2) For results of Table 1, the models are fine-tuned for 10 epochs with 128 batch size. The peak learning rate of the base-size and large-size models are set to 5e-5 and 3e-5, respectively. The input image resolution is 384×384. For ablation experiments, we fine-tune the models for 10 epochs with 128 batch size, and choose learning rates from {5e-5, 1e-4}. The input image resolution is 224×224. All the ablation results of NLVR2 are averaged over 3 runs.

2.3 Hyperparameters for Vision-Language Retrieval Fine-Tuning

COCO We fine-tune the base-size model for 20 epochs and large-size model for 10 epochs with 2048 batch size. The peak learning rate is 2e-5 for the base-size model and 1e-5 for the large-size model. The input image resolution is 384×384.

Flickr30K For results of Table 2, the base-size and large-size models are fine-tuned for 40 epochs with a batch size of 2048 and a peak learning rate of 1e-5. We use the fine-tuned model on COCO as the initialization. The input image resolution is 384×384. For all ablation experiments, we fine-tune the models for 10 epochs with 1024 batch size. The peak learning rate is set to 5e-5, and the input image resolution is 224×224.

References

[1] Junnan Li, Ramprasaath R. Selvaraju, Akhilesh Deepak Gotmare, Shafiq R. Joty, Caiming Xiong, and Steven C. H. Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *CoRR*, abs/2107.07651, 2021. URL `https://arxiv.org/abs/2107.07651`.

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

[3] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge. *IJCV*, 2015.

Models	ImageNet (acc@1)	ADE20K (mIoU)
ViT-Base	83.6	-
BEiT-Base	85.2	52.8
VLMo-Base	**85.5**	**53.4**

Table 3: Results on image classification and semantic segmentation.

[4] Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. Simvlm: Simple visual language model pretraining with weak supervision. *CoRR*, abs/2108.10904, 2021. URL https://arxiv.org/abs/2108.10904.

[5] Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ADE20K dataset. *Int. J. Comput. Vis.*, 127(3):302–321, 2019. doi: 10.1007/s11263-018-1140-0. URL https://doi.org/10.1007/s11263-018-1140-0.

[6] Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE international conference on computer vision*, pages 19–27, 2015.