Accelerating Vision-Language Pretraining with Free Language Modeling (Supplementary Materials)

Teng Wang[1,2], Yixiao Ge[3], Feng Zheng[1,5], Ran Cheng[1], Ying Shan[3], Xiaohu Qie[4], Ping Luo[2,6]
[1]Southern University of Science and Technology [2]The University of Hong Kong
[3]ARC Lab, [4]Tencent PCG [5]Peng Cheng Laboratory [6]Shanghai AI Laboratory
tengwang@connect.hku.hk {yixiaoge, yingsshan, tigerqie}@tencent.com
f.zheng@ieee.org ranchengcn@gmail.com pluo@cs.hku.hk

6. Supplementary Materials

6.1. Implementation Details

We list the hyperparameters for pretraining in Table 6. The implementation details for downstream tasks are described as follows.

Visual Question Answering (VQA). We follow [1] to consider VQA as a classification problem on 3129 most frequent answers. We input a single [CLS] token upon the reconstructor and regard its output representation as the multimodal features, followed by an MLP classifier to obtain the final classification probability. Following [2], during finetuning, the learning rates of the image encoder and bottom layers of the text transformer are 5e-6, and those for top layers of the text transformer and the reconstructor are 2.5e-5.

Natural Language for Visual Reasoning for Real (NLVR^2). The task aims to distinguish whether the natural language description is true given a pair of images. We follow [1] to consider the input triplet (a sentence and two images) as two image-text pairs. For each image-text pair, we obtain the [CLS] embedding from the reconstructor as the multimodal embeddings. The two embeddings are concatenated and input into an MLP for binary classification. Following [2], during fine-tuning, the learning rates of the image encoder and bottom layers of the text transformer and the reconstructor are set to 1e-5, and those for top layers of the text transformer are 5e-5.

Image Retrieval (IR) and Text Retrieval (TR). We use the image-text matching loss to finetune the pretrained model on the downstream retrieval datasets, *i.e.*, COCO and Flickr30K. During training, we construct random negative pairs by replacing the paired images with random images sampled from the dataset. An MLP is applied on the [CLS] embedding of the reconstructor for binary classification. The learning rates of the image encoder and bottom layers of the text transformer are 5e-6, and those for top layers and the reconstructor of the text transformer are 2.5e-5.

Image Captioning. Since the intermediate loss of our model considers an autoregressive generation process, the finetuning performance or zero-shot performance (shown in Sec. 6.2) on captioning datasets could be evaluated. For finetuning performance, we remove the reconstructor and finetune the model with unidirectional captioning loss. Note that we do not use beam search for simplicity. The learning rates of the image encoder and bottom layers of the text transformer are set to 3e-6, and those for top layers of the text transformer and the reconstructor are 1.5e-5.

6.2. Additional Analysis

More Evidence of the Motivation. In Table 7, our key motivation – limited prediction rate impedes convergence speed – is justified in wider environments with different VLP structures and pretraining data, *i.e.*, ViLT (single-encoder structure) and RoBERTa (pretrained on text-only datasets). A consistent trend is found that lower prediction rates gain higher MLM losses, verifying that such motivation is reasonable among different structures and datasets, even in text-only pretraining.

MLM with Varied Masking Ratios. We explore how much acceleration the MLM-based methods could achieve with a larger mask ratio. As shown in Table 8, when increasing the mask ratio from 0.2 to 0.8, a mask ratio of 0.6 achieves the best performance within the 30k steps. However, when the training steps grow after 50k steps, a mask rate of 0.4 achieves the best. Compared with 0.6, MLM with a 0.8 mask ratio shows slower convergence, probably caused by that larger corruption rate increasing the learning difficulty. We conclude that for MLM, the corruption rate and prediction rate are tied-up by the mask ratio, and a proper corruption rate is achieved at the cost of a large portion of output tokens being excluded from prediction loss.

	Ours$_{\text{BASE}}$	Ours$_{\text{LARGE}}$
patch size	32/16	14
image size	288×288	336×336
learning rate	4e-4	4e-4
learning rate (pretrained layers)	8e-5	8e-5
warmup rate	0.05	0.05
training steps	30k	30k

(a) 4M data

	Ours$_{\text{BASE}}$	Ours$_{\text{LARGE}}$
patch size	32/16	14
image size	288×288	224×224
learning rate	4e-4	4e-4
learning rate (pretrained layers)	8e-5	8e-5
warmup rate	0.05	0.05
training steps	30k	100k

(b) 13M data

Table 6. Hyper-parameters for pretraining.

Model	r_{corr}	r_{pred}	MLM loss	
			50% steps	100% steps
ViLT	40%	20%	2.161	2.044
ViLT	40%	40%	1.872	1.769
ViLT	15%	7.5%	1.808	1.655
ViLT	15%	15%	1.699	1.574
RoBERTa	40%	20%	3.857	3.501
RoBERTa	40%	40%	3.641	3.371

Table 7. Varying prediction and corruption rates. For ViLT [3], we follow the official recipe with 25k steps. For RoBERTa [4], we use an efficient recipe with 23k steps from [5].

Method	Mask Ratio	Training Steps					
		10k	20k	30k	50k	80k	100k
MLM	0.2	51.07	74.38	76.83	77.65	78.20	78.21
MLM	0.4	51.69	76.14	77.59	**78.46**	**78.64**	**78.30**
MLM	0.6	**62.62**	**76.69**	**78.01**	77.97	78.33	77.91
MLM	0.8	51.07	76.32	77.24	78.04	77.78	78.20

Table 8. NLVR^2 performance of different mask ratios in MLM. All models are trained with a maximum of 100k steps on 4M data.

PrefixLM with Varied r_{pred} and r_{corr}. Similarly to MLM, the corruption rate and prediction rate in PrefixLM are tied-up in nature. We found all experiments have a similar converge rate but much different converged performance. Table 9 shows the best result is achieved with $r_{\text{pred}} = 2 \cdot r_{\text{corr}} = 75\%$. Lower r_{corr} results in easier tasks and lower representability, while a much larger one may cause learning collapse.

Zero-shot Captioning Performance. The proposed FLM objectives include two parts, a reconstruction loss for solving corruption-prediction tasks with bidirectional contexts, and an intermediate loss that supervises the model and focuses more on temporal relationships with unidirectional context. After pretraining, we could directly test the zero-shot captioning performance without further finetuning on target datasets. The zero-shot performance of different pretraining objectives is shown in Table 10. While

Method	r_{corr}	r_{pred}	NLVR^2
PrefixLM	12.5%	25%	75.00
PrefixLM	25.0%	50%	76.17
PrefixLM	37.5%	75%	76.78
PrefixLM	50.0%	100%	76.29

Table 9. Varying r_{corr} and r_{pred} of PrefixLM. It is achieved by modifying the distribution of prefix length. All models are trained with a maximum of 30k steps on 4M data.

Method	Zero-shot Captioning			Finetuned Captioning		
	B@4	M	C	B@4	M	C
AR	**24.9**	**21.6**	**80.3**	35.70	28.86	120.6
PrefixLM	22.6	20.2	73.3	35.50	28.79	119.4
FLM	20.7	19.6	70.3	**36.68**	**29.17**	**123.0**

Table 10. Image captioning performance of different pretraining objectives on COCO. B@4, M, C are short for BLEU@4, METEOR, CIDEr, respectively.

MLM-based methods can not be directly used in captioning tasks, FLM achieves reasonable zero-shot captioning performance. However, for finetuned captioning performance, FLM achieves better performance than AR/PrefixLM. We conjecture that the FLM objectives could capture more generalizable features than AR/PrefixLM for captioning after finetuning.

References

[1] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. UNITER: Universal image-text representation learning. In *European Conference on Computer Vision (ECCV)*, 2020. 1

[2] Zi-Yi Dou, Yichong Xu, Zhe Gan, Jianfeng Wang, Shuohang Wang, Lijuan Wang, Chenguang Zhu, Pengchuan Zhang, Lu Yuan, Nanyun Peng, et al. An empirical study of training end-to-end vision-and-language transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18166–18176, 2022. 1

[3] Wonjae Kim, Bokyung Son, and Ildoo Kim. ViLT: Vision-and-language transformer without convolution or region su-

pervision. In *International Conference on Machine Learning (ICML)*, 2021. 2

[4] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A robustly optimized bert pretraining approach. *arXiv preprint*, 2019. 2

[5] Alexander Wettig, Tianyu Gao, Zexuan Zhong, and Danqi Chen. Should you mask 15% in masked language modeling? *arXiv preprint arXiv:2202.08005*, 2022. 2