mc-BEiT: Multi-choice Discretization for Image BERT Pre-training (Appendix)

Xiaotong Li[1],Yixiao Ge[2],Kun Yi[2],Zixuan Hu[1],Ying Shan[2],Ling-Yu Duan[1,3,*]

[1]Peking University, Beijing, China [2]ARC Lab, Tencent PCG
[3]Peng Cheng Laboratory, Shenzhen, China
lixiaotong@stu.pku.edu.cn, {hzxuan, lingyu}@pku.edu.cn,
{yixiaoge, yingsshan, kunyi}@tencent.com
[*]Corresponding Author

A Implementation Details

In the appendix, we provide the specific hyper-parameters of the experiments in our paper, including pre-training on ImageNet-1K and fine-tuning on different downstream tasks.

A.1 Configuration for pre-training

The vision Transformers are pre-trained on the large-scale dataset ImageNet-1K [5] and the configurations are summarized in Tab. 1. The implementation of the vision Transformers, *i.e.,* ViT-Base/16 and ViT-Large/16, follows [10] for fair comparisons and the training recipe is based on BEiT [1].

A.2 Configuration for fine-tuning

Classification task on ImageNet-1K For the classification task, the fully-connected layer is employed as the classifier after the average pooling of the feature embeddings. The fine-tuning configurations on ImageNet-1K for different backbone architectures are listed in Tab. 2.

Object detection and instance segmentation We adopt the implementation of [9, 7] to verify our performances of object detection and instance segmentation on COCO. Tab. 3 summarizes the configurations for fine-tuning on COCO. The training recipe of models with intermediate fine-tuning is the same as the pre-training only version. ViT-B [6] is adopted as the backbone and Mask-RCNN [8] is used as the task head.

Besides, we also provide another experiment result following the implementation in iBOT [13] in Tab. 4. Because these experiments are not conducted on BEiT, we conduct the experiments following iBOT [13] and the results of BEiT [1] are based on our re-implementation. In order to adapt to the multi-scale strategy, we use absolute position embedding and interpolate it for different image

Table 1: Configurations for pre-training.

Configuration	ViT-Base/16	ViT-Large/16
Layers	12	24
Hidden size	768	1024
FFN inner hidden size	3076	4096
Attention heads	12	16
Attention head size	64	
Patch size	16×16	
Training epochs	800	
Batch size	2048	
Adam ϵ	1e-8	
Adam β	(0.9, 0.98)	
Peak learning rate	1.5e-3	
Minimal learning rate	1e-5	
Learning rate schedule	Cosine	
Warmup epochs	10	
Gradient clipping	3.0	1.0
Dropout	None	
Stoch. depth	0.1	
Weight decay	0.05	
Data Augment	RandomResizeAndCrop	
Input resolution	224×224	

resolutions. ViT-B [6] is adopted as the backbone and Cascaded Mask-RCNN [8, 2] is used as the task head.

Table 3: Configurations for fine-tuning on COCO.

Configuration	ViT-Base/16
Fine-tuning epochs	25
Peaking learning rate	8e-5
Learning rate decay	cosine
Adam ϵ	1e-8
Adam β	(0.9, 0.999)
Dropout	None
Stoch. depth	0.1
Weight decay	0.1
Batch size	64
Input size	1024×1024
Position embedding	Abs. + Rel.
Augmentation	LSJ(0.1, 2.0)

Table 2: Configurations for fine-tuning on ImageNet-1K.

Configuration	ViT-Base/16	ViT-Large/16
Peak learning rate	{2e-3,3e-3,4e-3,5e-3}	
Fine-tuning epochs	100	50
Batch size	1024	
Warmup epochs	20	5
Layer-wise learning rate decay	0.65	0.75
Adam ϵ	1e-8	
Adam β	(0.9, 0.999)	
Minimal learning rate	1e-6	
Learning rate schedule	Cosine	
Repeated Aug	None	
Weight decay	0.05	
Label smoothing	0.1	
Stoch. depth	0.1	
Dropout	None	
Gradient clipping	None	
Erasing prob.	0.25	
Input resolution	224×224	
Rand Augment	9/0.5	
Mixup prob.	0.8	
Cutmix prob.	1.0	
Color jitter	0.4	

Semantic segmentation on ADE20K: For the semantic segmentation experiments on ADE20K [12], we follow the implementation of BEiT [1] and adopt UperNet [11] as the task layer. ViT-B [6] is adopted as the default backbone and UPerNet [11] is used as the task head. Tab. 5 summarizes the configurations for fine-tuning on ADE20k. Because the pre-training process does not introduce the instance discrimination, the performance can be further improved after intermediate fine-tuning on ImageNet-1K according to BEiT [1]. we also evaluate the performances after intermediate fine-tuning, where the pre-trained models have been fine-tuned on ImageNet-1K. For the models with intermediate fine-tuning, the training recipe is the same as the pre-training only version.

Table 4: We provide another experiment results of object detection and instance segmentation on COCO following the implementation of iBOT[13]. Intermediate fine-tuning denotes the model is further fine-tuned on ImageNet-1K. Cascaded Mask R-CNN and 1× training schedule are adopted.

Method	Reference	Object Det. AP^b	Instance Seg. AP^m
Supervised [10]	ICML 2021	47.9	42.9
MoCo v3 [4]	CVPR 2021	47.9	42.7
DINO [3]	ICCV 2021	50.1	43.4
iBOT [13]	ICLR 2022	51.2	44.2
BEiT [1]	ICLR 2022	49.6	42.8
Ours	this paper	50.1	43.1
+Intermediate Fine-tuning			
BEiT [1]	ICLR 2022	50.7	43.8
Ours	this paper	**51.2**	**44.3**

Table 5: Configurations for fine-tuning on ADE20k.

Configuration	ViT-Base/16
Peaking learning rate	8e-5
Fine-tuning steps	160000
Batch size	16
Adam ϵ	1e-8
Adam β	(0.9, 0.999)
Layer-wise learning rate decay	0.9
Minimal learning rate	0
Learning rate schedule	Linear
Warmup steps	1500
Dropout	None
Stoch. depth	0.1
Weight decay	0.05
Input resolution	512×512
Position embedding	Relative
Position embedding interpolate	Bilinear

References

1. Bao, H., Dong, L., Piao, S., Wei, F.: BEit: BERT pre-training of image transformers. In: ICLR (2022)
2. Cai, Z., Vasconcelos, N.: Cascade r-cnn: high quality object detection and instance segmentation. TPAMI pp. 1483–1498 (2019)
3. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: ICCV (2021)
4. Chen, X., Xie, S., He, K.: An empirical study of training self-supervised vision transformers. In: ICCV. pp. 9640–9649 (2021)
5. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: CVPR. pp. 248–255
6. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: ICLR (2021)
7. Fang, Y., Yang, S., Wang, S., Ge, Y., Shan, Y., Wang, X.: Unleashing vanilla vision transformer with masked image modeling for object detection. arXiv preprint arXiv:2204.02964 (2022)
8. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: ICCV. pp. 2980–2988 (2017)
9. Li, Y., Xie, S., Chen, X., Dollar, P., He, K., Girshick, R.: Benchmarking detection transfer learning with vision transformers. arXiv preprint arXiv:2111.11429 (2021)
10. Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jegou, H.: Training data-efficient image transformers and distillation through attention. In: ICML. pp. 10347–10357 (2021)
11. Xiao, T., Liu, Y., Zhou, B., Jiang, Y., Sun, J.: Unified perceptual parsing for scene understanding. In: ECCV (September 2018)
12. Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., Torralba, A.: Scene parsing through ade20k dataset. In: CVPR. pp. 5122–5130 (2017)
13. Zhou, J., Wei, C., Wang, H., Shen, W., Xie, C., Yuille, A., Kong, T.: Image BERT pre-training with online tokenizer. In: ICLR (2022)