TinyMIM: An Empirical Study of Distilling MIM Pre-trained Models Supplementary Material

Sucheng Ren   Fangyun Wei*   Zheng Zhang   Han Hu
Microsoft Research Asia

A. Hyper-parameters

Pre-training. All models are pre-trained under a 300-epoch schedule on ImageNet-1K [7] training set. We use a batch size of 4096 and a learning rate of lr=1.5e-4$\times$batchsize/256. We adopt a cosine decay schedule with a warm-up for 15 epochs. We adopt AdamW [6] optimizer with a weight decay of 0.05. We use random resized cropping random horizontal flipping, color jitter for student only. The input size is set to 224×224.

Fine-tuning. We transfer TinyMIM pre-trained models to ImageNet [7] image classification and ADE20K [10] semantic segmentation. For ImageNet, we use AdamW optimizer with weight decay of 0.05. For data augmentation, we follow the settings in MAE [1]. We fine-tune ViT-B for 100 epochs with a batch size of 1024, a learning rate of 2e-3, and a drop path rate of 0.1. We fine-tune ViT-S and ViT-T for 200 epochs with a batch size of 2048, a learning rate of 5e-3, and a drop path rate of 0.1. For ADE20K, we follow the same setting in MAE and adopt UperNet [9] as our framework with a TinyMIM pre-trained backbone. The input image resolution is 512×512 for training and evaluating. We use mIoU as the evaluation metric.

Besides, we evaluate the robustness of TinyMIM on various out-of-domain ImageNet datasets [2–4] which are generated by applying different perturbations on ImageNet, *e.g.* natural adversarial examples (ImageNet-A), semantic shift (ImageNet-R), common image corruptions (ImageNet-C). We report top-1 accuracy on ImageNet-A/R and mCE error on ImageNet-C (lower is better).

Hyper-parameters of ImageNet-1K Pre-training. See Table 1.

Hyper-parameters of ImageNet-1K Image Classification Fine-tuning. See Table 2. TinyMIM*-T retains the plain architecture and computation budget of ViT-T. We fine-tune TinyMIM* for 1000 epochs with DeiT-style [8] knowledge distillation on ImageNet-1K. Following MobileNetV3 [5], an extra fully connected layer is placed before the classification layer to increase the feature dimension from 192 to 1280. The head number is set to 12 instead of the default 3.

Hyper-parameters for ADE20K Semantic Segmentation Fine-tuning. See Table 3.

Hyperparameter	**ViT-T**	**ViT-S**	**ViT-B**
Layers	12		
Hidden size	192	384	768
FFN inner hidden size	768	1536	3072
Attention heads	3	6	12
Patch size	16×16		
Pre-training epochs	100/300		
Batch size	4096		
Adam ϵ	1e-8		
Adam β	(0.9, 0.95)		
Peak learning rate	2.4e-3		
Minimal learning rate	1e-5		
Learning rate schedule	Cosine		
Warmup epochs	5/15		
Stochastic depth	0.1		
Dropout	✗		
Weight decay	0.05		
Data augment	RandomResizeAndCrop		
Input resolution	224×224		
Color jitter (student only)	0.4		

Table 1. Hyper-parameters of ImageNet-1K Pre-training.

[*] Corresponding author: fawe@microsoft.com.

Hyperparameter	ViT-T	ViT-S	ViT-B
Peak learning rate	5e-3	5e-3	2e-3
Fine-tuning epochs	200	200	100
Warmup epochs		5	
Layer-wise learning rate decay	0.65	0.65	0.65
Batch size	2048	2048	1024
Adam ϵ		1e-8	
Adam β		(0.9, 0.95)	
Learning rate schedule		Cosine	
Stochastic depth	0.1	0.1	0.1/0.2*
Weight decay		0.05	
Label smoothing ε		0.1	
Dropout		✗	
Gradient clipping		✗	
Erasing		0.25	
Input resolution		224 × 224	
Rand augment		9/0.5	
Mixup		0.8	
Cutmix		1.0	

Table 2. Hyper-parameters of ImageNet-1K image classification fine-tuning. * indicates that we use 0.1 and 0.2 for 100-epoch and 300-epoch pre-trained models, respectively.

Hyperparameter	ViT-S	ViT-B
Input resolution	512 × 512	
Peak learning rate	1e-4	
Fine-tuning steps	160K	
Batch size	16	
Adam ϵ	1e-8	
Adam β	(0.9, 0.999)	
Layer-wise learning rate decay	$\{0.65, 0.75, 0.8\}$	
Minimal learning rate	0	
Learning rate schedule	Linear	
Warmup steps	1500	
Dropout	✗	
Stochastic depth	0.1	
Weight decay	0.05	

Table 3. Hyper-parameters of ADE20K semantic segmentation fine-tuning.

References

[1] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, 2022. 1

[2] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, and Justin Gilmer. The many faces of robustness: A critical analysis of out-of-distribution generalization. *ICCV*, 2021. 1

[3] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *ICLR*, 2019. 1

[4] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. *CVPR*, 2021. 1

[5] Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, et al. Searching for mobilenetv3. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1314–1324, 2019. 1

[6] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. 1

[7] 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. 1

[8] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. *preprint arXiv:2012.12877*, 2020. 1

[9] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In *ECCV*, 2018. 1

[10] 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. 1