SparseMAE: Sparse Training Meets Masked Autoencoders

Aojun Zhou[1] Yang Li[2] Zipeng Qin[1] Jianbo Liu[1] Junting Pan[1]
Renrui Zhang[13] Rui Zhao[2] Peng Gao[3] Hongsheng Li[1]
[1]The Chinese University of Hong Kong
[2]SenseTime Research [3]Shanghai AI Lab

Appendix

A.Hyper-parameters

Hyper-parameters of ImageNet-1K Pretraining. See Tab 1.

Hyper-parameters of ImageNet-1K Image Classification Finetuning. See Tab 2. We fine-tune SparseMAE-T/S for 200 epochs following TinyMIM and G2SD on ImageNet-1K.

Hyper-parameters for ADE20K Semantic Segmentation Finetuning. See Tab 3.

Table 1: Hyper-parameters of ImageNet-1K pretraining.

Hyperparameter	**SparseMAE-T**	**-S**
Layers	12	
Mask ratio	25%	25%
Sparse pattern	2:32	4:32
Patch size	16×16	
Pre-training epochs	400	
Batch size	4096	
Adam ϵ	1e-8	
Adam β	(0.9, 0.999)	
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	

Table 2: Hyper-parameters of ImageNet-1K image classification finetuning.

Hyperparameter	**SparseMAE-T**	**-S**
Peak learning rate	5e-3	
Fine-tuning epochs	200	
Warmup epochs	5	
Layer-wise learning rate decay	0.65	
Batch size	2048	
Adam ϵ	1e-8	
Adam β	(0.9, 0.999)	
Minimal learning rate	1e-6	
Learning rate schedule	Cosine	
Stochastic depth	0.1	
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 3: Hyper-parameters of ADE20K semantic segmentation finetuning.

Hyperparameter	**SparseMAE-T SparseMAE-S**
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