Improving CLIP Fine-tuning Performance

——— Supplementary Material ———

Yixuan Wei[1], Han Hu[2]*, Zhenda Xie[1], Ze Liu[3], Zheng Zhang[2], Yue Cao[2],
Jianmin Bao[2], Dong Chen[2], Baining Guo[2]
[1]Tsinghua University [2]Microsoft Research Asia [3]USTC
{t-yixuanwei,hanhu,t-zhxie,t-liuze,zhez,yuecao,jianbao,doch,bainguo}@microsoft.com

A. Additional training costs

We briefly discussed this in limitations of the main body: a **3%** additional training cost is needed to improve the fine-tuning performance of CLIP through our method. Here is how we get this number. Most of our results are obtained using ImageNet-1k for 300 epochs, which involved training on an additional 384M images (300×1.28M). During CLIP pre-training, it is trained on WIT-400M dataset for 32 epochs, equal to 12.8B image instances (we omit the pre-training cost of the text encoder, since it was not as heavy as the image encoder). Therefore, the 3% additional cost is calculated as **384M/12.8B**, which is affordable considering the performance boost it provides during fine-tuning. In addition, unlike CLIP pre-training which requires a large number of GPUs to achieve a sufficient batch size (the original CLIP model was trained with a batch size of 32768 and 256 V100 GPUs), our method only requires a small batch size of 2048 and 8 V100 GPUs, making it accessible to most labs and groups.

B. Results of FD-MAE

Similar to **FD**-CLIP, we took the MAE ViT-B as the teacher and distilled it for 300 epochs on ImageNet-1k. The results are listed in the table below. The **FD**-MAE performed similar to its teacher on most tasks, verifying our observations that the gain of our method is largely from a token-level task which is already used in MAE pre-training.

Table 1: Results of **FD**-MAE.

Method	IN-1K %	ADE20K mIoU	COCO		NYUv2 RMSE (↓)
			AP_{box}	AP_{mask}	
MAE	83.6	48.1	46.5	40.9	0.383
FD-MAE	83.4	47.9	46.7	41.2	0.364
Δ	**↓0.2**	**↓0.2**	**↑0.2**	**↑0.3**	**↓0.019**

*Corresponding Author. The work is done when Yixuan Wei, Zhenda Xie, and Ze Liu are interns at Microsoft Research Asia.

C. Longer epoch for masked versions.

The masked version may reduce the additional **3%** cost to be even smaller. However, honestly, by using longer epochs in a masked version, we did not find any gains over our full version, as shown in the table below. One possible improving direction is to use some advanced masking methods [4]. We will leave this as future work.

Table 2: Longer training epoch with masked input shows inferior performance.

Method	GPU Time	IN-1K %	ADE20K mIoU	COCO		NYUv2 RMSE (↓)
				AP_{box}	AP_{mask}	
25% input + 100ep	96.4h	83.1	48.8	45.1	39.8	0.379
25% input + 200ep	192.8h	83.9	49.7	46.4	40.9	0.366
25% input + 400ep	385.6h	**84.4**	51.1	47.2	41.5	0.367
Full input + 100ep	170.7h	**84.4**	**51.8**	**47.9**	**42.2**	**0.350**

D. *Shared* RPB enhanced the diversity of heads

We diagnose the effects of using different position encoding configurations during feature distillation on CLIP ViT-B/16, including APE, *non-shared* RPB, and *shared* RPB (the default setting). Their average attention distances per head are visualized in Fig. 1. Compared to the models that use APE and *non-shared* RPB, the *shared* RPB can diversify the attention distances of heads a bit more, especially for the deeper layers, which may cause its slightly better fine-tuning accuracy, i.e., +0.4∼0.5% top-1 accuracy on ImageNet-1K classification.

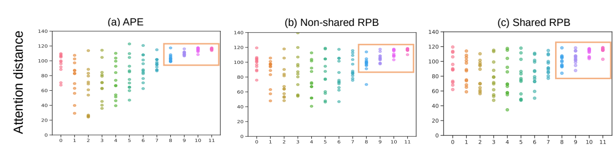

Figure 1: Comparison of average attention distances per head after distillation for different position encoding configurations.

E. Full average attention maps of MAE, CLIP, and FD-CLIP

In the main body, we have visualized the average attention maps of 5 representative layers for MAE, CLIP, and **FD**-CLIP. Here, we supplement with the average attention maps of all layers (Layer 0-11 are visualized from top-left to bottom-right): MAE in Fig. 2, CLIP in Fig. 3 and **FD**-CLIP in Fig. 4. In the visualization, the image patches (total 196) are indexed starting from top-left to bottom-right. From these visualizations, we can draw a conclusion that aligns with our observation in the main body, *i.e.*, the model after distillation learns better inductive bias of translational invariance and locality prior, showing more *diagonal* and less *vertical-bar* attention patterns.

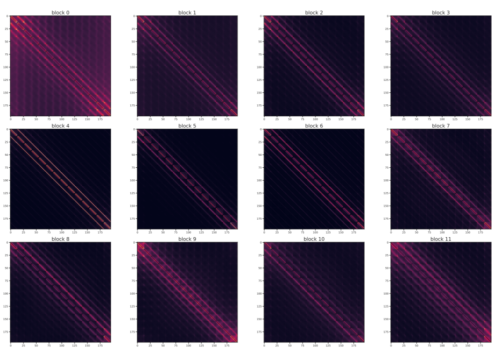

Figure 2: All 12 layers' average attention maps on MAE.

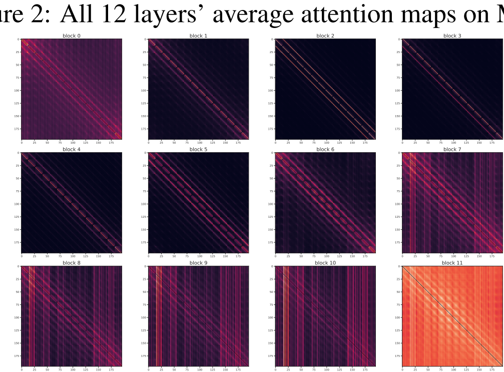

Figure 3: All 12 layers' average attention maps on CLIP.

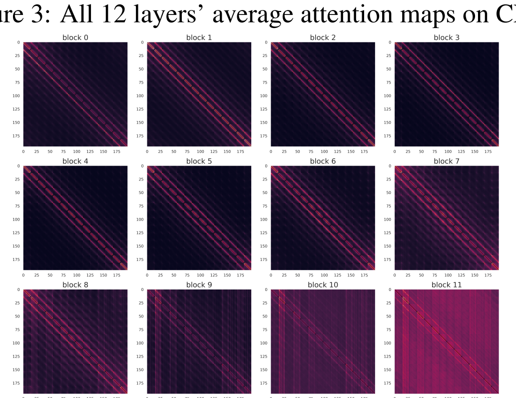

Figure 4: All 12 layers' average attention maps on **FD**-CLIP.

F. Further boosting ImageNet-1K classification with advanced tricks [2]

After our previous submission to CVPR 2023, a sophisticated and detailed fine-tuning recipe [2] on ImageNet-1K classification for CLIP is proposed. With careful hyper-parameters tuning, such as learning rate, stochastic depth rate, data augmentation strength, and training epochs, and introducing advanced techniques, performance on ImageNet-1K classification is pushed to 85.7% top-1 accuracy for CLIP ViT-B/16. Inspired by their findings, we also carefully fine-tuned our models with new recipes, as shown in Tab. 3. **FD**-CLIP still earns clear performance gains on both base- and large-size models under sophisticated recipes.

Also note the new recipe [2] only effects for image classification performance. **Our method still shows significant advantages on dense prediction tasks including detection, segmentation and depth estimation, with careful hyper-parameter fine-tuning.**

Table 3: Boosting feature distillation on ImageNet-1K with advanced fine-tuning recipes. C. means COCO.

Method	B/16$_{224}$					L/14$_{224}$
	IN-1K	ADE20K	C. AP$_{box}$	C. AP$_{mask}$	NYUv2 (↓)	IN-1K
Arxiv22 [2]	85.7	49.5	45.0	39.8	0.416	88.0
FD-CLIP	**85.9** (+0.2)	**51.7** (+2.2)	**48.2** (+3.2)	**42.5** (+2.7)	**0.352** (-0.064)	**88.4** (+0.4)

G. Hyperparameters for Feature Distillation

Table 4 lists the hyperparameters used in the feature distillation method.

H. Hyperparameters for Fine-tuning

Fine-tuning on ImageNet-1K classification. Table 5 lists the hyperparameters used for fine-tuning on imagenet-1K.

Fine-tuning on COCO object detection and instance segmentation. We implement the Mask R-CNN framework following MMDetection [1]. The batch size is 16, the learning rate is 2e-4, and the layer-wise decay rate is 0.75. Following the common practice, we decay the learning rate by 10× at epochs 9 and 11.

Fine-tuning on NYUv2 depth estimation. The NYUv2 dataset includes an official training split (24K images) and official testing split with 654 images from 215 indoor scenes. The head of the depth estimation and the data augmentations are following [3]. And we also average the prediction of the two square windows in testing.

References

[1] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu,

Table 4: Hyperparameters for feature distillation on ImageNet-1K.

Hyperparameters	**Base Size**	**Large Size**
Patch size	16×16	14×14
Layers	12	24
Hidden size	768	1024
FFN inner hidden size	3072	4096
Attention heads	12	16
Attention head size	64	
Training epochs	300	
Batch size	2048	
Adam ϵ	1e-8	
Adam β	(0.9, 0.999)	
Peak learning rate	1.2e-3	
Minimal learning rate	2e-5	
Learning rate schedule	Cosine	
Warmup epochs	10	
Gradient clipping	3.0	
Dropout	✗	
Weight decay	0.05	
Stoch. depth	{0.1,0.2,0.3}	0.3
Data Augment	RandomResizeAndCrop 0.08-1	
Input resolution	224×224	

Table 5: Hyperparameters for fine-tuning on ImageNet-1K.

Hyperparameters	**Base Size**	**Large Size**
Peak learning rate	{5e-3, 6e-3}	1e-3
Fine-tuning epochs	100	50
Warmup epochs	20	5
Layer-wise learning rate decay	{0.6, 0.65}	0.75
Batch size	2048	
Adam ϵ	1e-8	
Adam β	(0.9, 0.999)	
Minimal learning rate	2e-6	
Learning rate schedule	Cosine	
Repeated Aug	✗	
Weight decay	0.05	
Label smoothing ε	0.1	
Stoch. depth	{0.1,0.2,0.3}	0.4
Dropout	✗	
Gradient clipping	5.0	
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	

Jiarui Xu, et al. Mmdetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*, 2019. 2

[2] Xiaoyi Dong, Jianmin Bao, Ting Zhang, Dongdong Chen, Shuyang Gu, Weiming Zhang, Lu Yuan, Dong Chen, Fang Wen, and Nenghai Yu. Clip itself is a strong fine-tuner: Achieving 85.7% and 88.0% top-1 accuracy with vit-b and vit-l on imagenet. *arXiv preprint arXiv:2212.06138*, 2022. 2

[3] Zhenda Xie, Zigang Geng, Jingcheng Hu, Zheng Zhang, Han Hu, and Yue Cao. Revealing the dark secrets of masked image modeling. *arXiv preprint arXiv:2205.13543*, 2022. 2

[4] Yifan Yang, Weiquan Huang, Yixuan Wei, Houwen Peng, Xinyang Jiang, Huiqiang Jiang, Fangyun Wei, Yin Wang, Han Hu, Lili Qiu, et al. Attentive mask clip. *arXiv preprint arXiv:2212.08653*, 2022. 1