Appendix for "DisCo-CLIP: A Distributed Contrastive Loss for Memory Efficient CLIP Training"

Yihao Chen, Xianbiao Qi, Jianan Wang, Lei Zhang
International Digital Economy Academy (IDEA), Shenzhen, Guangdong, China.
{chenyihao, qixianbiao, wangjianan, leizhang}@idea.edu.cn

1. Appendix

1.1. Training Curves

To facilitate the readers to observe the training process, we show the curves of the training loss and the zero-shot top-1 ImageNet classification in Fig. 1. The base model is ViT-B/32, the used batch size is 65,536. The model is trained 32 epochs on LAION-400M.

1.2. Zero-shot Classification Details

We follow CLIP [3] for the zero-shot classification. We use the same 80 prompts as in CLIP. During evaluation, We first resize the image to the short side 224 according to the image ratio and then centrally crop a 224×224 image.

Dataset	Classes	Number of Test Images
INet [1]	1,000	50,000
INet-v2 [4]	1,000	10,000
INet-R [2]	1,000	30,000
INet-S [5]	1,000	50,889

Table 1. Information of the evaluation data sets used in our paper.

Our used evaluation datasets include INet [1], INet-v2 [4], INet-R [2] and INet-S [5]. We show some basic information in Tab. 1. More information can be found in Timm[1].

1.3. More Experiments

We further evaluate the performance of DisCo-CLIP under different batch sizes. The experiments are conducted on LAION-100M subset, the models are based on ViT-B/32, and the number of training epochs is 16. The results are shown in Tab. 2. We can see that from Tab. 2 larger batch size always brings in performance gain. Using DisCo-CLIP, it takes around 2 days to finish training of 16 epochs of LAION-100M with batch size 32,768 on a cluster with only 8 A100 GPUs.

[1] https://github.com/rwightman/pytorch-image-models

References

[1] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255. IEEE, 2009. 1, 2

[2] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8340–8349, 2021. 1, 2

[3] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 1

[4] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *International Conference on Machine Learning*, pages 5389–5400. PMLR, 2019. 1, 2

[5] Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. *Advances in Neural Information Processing Systems*, 32, 2019. 1, 2

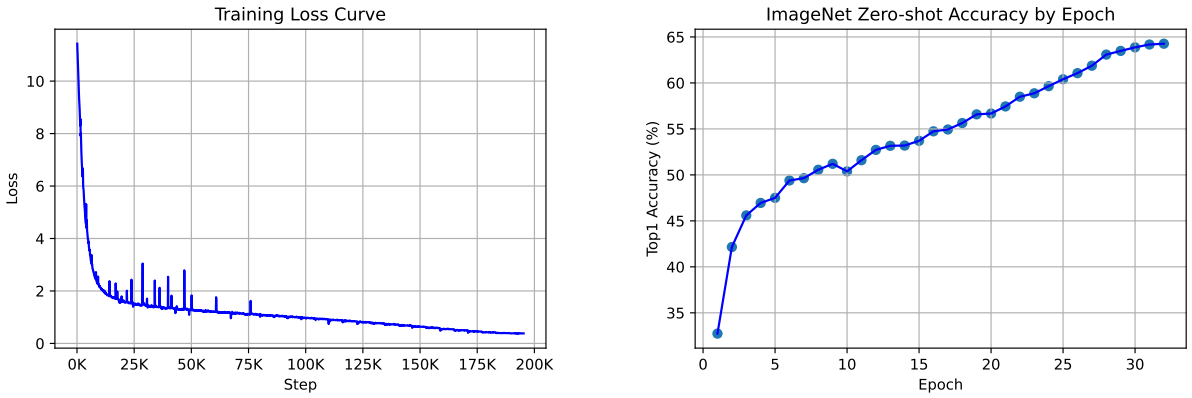

Figure 1. Curve of the training loss on the left and curve of the zero-shot top-1 ImageNet classification accuracy (%) on the right. The base model is ViT-B/32, the used batch size is 65,536. The model is trained 32 epochs on LAION-400M.

Model	Datasets	Epochs	Steps	Batch Size	INet [1]	INet-V2 [4]	INet-R [2]	INet-S [5]
DisCo-CLIP	LAION-100M*	16	200 K	8,192	48.76	41.28	56.67	36.65
DisCo-CLIP	LAION-100M*	16	100 K	16,384	50.95	43.07	59.24	38.74
DisCo-CLIP	LAION-100M*	16	50 K	32,768	51.64	43.85	60.07	39.25
DisCo-CLIP	LAION-100M*	16	25 K	65,536	**51.91**	**44.19**	**60.52**	**39.76**

Table 2. Performance evaluation of DisCo-CLIP under different batch size. We report top-1 zero-shot classification accuracy (%) on several data sets. All models are based on ViT-B/32. Our LAION-100M* is a 100M subset of LAION-2B.