Supplementary Material: Mutual Conditional Probability for Self-Supervised Learning

Takumi Kobayashi[1,2]
takumi.kobayashi@aist.go.jp

[1] National Institute of Advanced Industrial Science and Technology
Tsukuba, Japan

[2] University of Tsukuba
Tsukuba, Japan

1 Experimental Setting

In the experiments (Sec. 3), we apply the setting shown in Table 1. We utilize the standard data augmentation process [1, 4] to produce 4 views of 160×160 image sizes for our method.

Architecture. The network (Fig. 2) contains the projection φ and probe ψ heads which are implemented by two-layer MLP composed of linear projection, batch normalization and ReLU; the hidden and output dimensions are also detailed in Table 1. As described in Sec. 2.2, the backbone ϕ and projection φ are subject to exponential moving averaging (EMA), so-called momentum encoder [5], to slowly update samples as in [4] via

$$\theta^{t+1} = (1-\tau)\theta + \tau\theta^t, \quad \tau = 1 - (1-\tau_0)\frac{\cos\left(\frac{\pi t}{t_{\text{end}}}\right)+1}{2}, \tag{1}$$

where θ^t is a model parameter of ϕ and φ at the t-th step, t_{end} indicates the total number of training steps and τ_0 is a base EMA factor.

Optimizer. We apply Adam [6] optimizer with 500-step linear warm-up to the datasets other than ImageNet for which SGD with momentum of 0.9 is used without warm-up. In the Adam optimizer, the learning rate is dropped by the factor of 0.2 at the 50 and 25 epochs *before* the last epoch. The learning rate of SGD is decayed in a cosine schedule [9].

Linear evaluation. Linear classifiers are trained on labeled samples by applying SGD with 0.9 momentum to ImageNet with the initial learning rate of 0.01 decayed by 0.1 at the 15 and 30 epochs, while we apply to the other datasets Adam [6] optimizer with exponentially decayed learning rate from 10^{-2} to 10^{-6}.

Computation resource. We implemented models by PyTorch on 4 NVIDIA V100 GPUs.

In summary, Table 1 details the datasets as well as the experimental settings used in Sec. 3; we basically follow the protocol of [7] for ImageNet and that of [3] for the other datasets.

Table 1: Detailed settings. On SUN-397, we evaluate models on split1 while the given training/test splits are used for the other datasets.

	ImageNet [7]	ImageNet-100 [10]	SUN397 [11]	ImageNet-LT [8]
Dataset stats.				
Training sample	1,281,167	126,689	76,240	115,846
Labeled sample	1,281,167	126,689	19,850	115,846
Test sample	50,000	5,000	19,850	50,000
Classes	1,000	100	397	1,000
Architecture				
Backbone	ResNet50	ResNet18	ResNet18	ResNet18
MLP output dim.	256	128	128	128
MLP hidden dim.	4096	1024	1024	1024
EMA factor τ_0	0.99	0.99	0.99	0.99
SSL training				
Optimizer	SGD	Adam	Adam	Adam
Learning rate	0.1	0.002	0.001	0.001
Schedule	Cosine	Drop	Drop	Drop
Weight Decay	10^{-4}	10^{-6}	10^{-6}	10^{-6}
Epochs	100	240	240	240
Batch size	256	512	512	512
Linear eval. training				
Optimizer	SGD	Adam	Adam	Adam
Batch Size	256	1000	1000	1000
Learning rate	0.01	0.01	0.01	0.01
Schedule	Drop	Exp.	Exp.	Exp.
Weight Decay	10^{-4}	$5 \cdot 10^{-6}$	$5 \cdot 10^{-6}$	$5 \cdot 10^{-6}$
Epochs	40	500	500	500

References

[1] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *CVPR*, 2021.

[2] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255, 2009.

[3] Aleksandr Ermolov, Aliaksandr Siarohin, Enver Sangineto, and Nicu Sebe. Whitening for self-supervised representation learning. In *ICML*, 2021.

[4] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. Bootstrap your own latent: A new approach to self-supervised learning. In *NeurIPS*, 2020.

[5] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and RossGirshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 2020.

[6] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015.

[7] Soroush Abbasi Koohpayegani, Ajinkya Tejankar, and Hamed Pirsiavash. Mean shift for self-supervised learning. In *ICCV*, pages 10326–10335, 2021.

[8] Ziwei Liu, Zhongqi Miao, Xiaohang Zhan, Jiayun Wang, Boqing Gong, and Stella X Yu. Large-scale long-tailed recognition in an open world. In *CVPR*, pages 2537–2546, 2019.

[9] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. In *International Conference on Learning Representations (ICLR)*, 2017.

[10] Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *ICML*, 2020.

[11] Jianxiong Xiao, James Hays, Krista A. Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *CVPR*, 2010.