Rethinking Rotation in Self-Supervised Contrastive Learning: Adaptive Positive or Negative Data Augmentation

Atsuyuki Miyai[1] Qing Yu[1] Daiki Ikami[2] Go Irie[2] Kiyoharu Aizawa[1]
[1]The University of Tokyo [2]NTT Corporation, Japan
{miyai,yu,aizawa}@hal.t.u-tokyo.ac.jp daiki-ikami@go.tuat.ac.jp goirie@ieee.org

1. Experimental details

Detailed experimental setups of SimCLR, MoCo v2 and BYOL are given below.

1. **SimCLR.** We use ResNet-18* and ResNet-50* for CIFAR-100 and use ResNet-18 and ResNet-50 for Tiny ImageNet. These networks are followed by the two-layer multilayer perceptron (MLP) projection head (output dimensions are 128). We set τ to 0.5 in all experiments. For data augmentations, we adopt basic augmentations proposed by Chen et al. [1]: namely, inception crop, horizontal flip, color jitter, and grayscale. On CIFAR-100, models are trained for up to 300 epochs with a batch size of 128. On Tiny ImageNet, models are trained for 200 epochs with a batch size of 256. For optimization, we train under LARS optimizer [4] with a weight decay of 1e-6 and a momentum with 0.9. An initial learning rate is 0.20. For the learning rate scheduling, we use linear warmup [2] for early 10 epochs and decay with cosine decay schedule without a restart [3]. For evaluation, we train a linear classifier for 90 epochs with a batch size of 128 using stochastic gradient descent with a momentum of 0.9 in all experiments. The learning rate starts at 0.1 and is dropped by a factor of 10 at 60%, 75%, and 90% of the training progress. We conducted the training on a single Nvidia V100 GPU.

2. **MoCo v2.** We use ResNet-18* and ResNet-50* for CIFAR-100 and ResNet-18, ResNet-50 and ResNet-18* for Tiny ImageNet. These networks are followed by the two-layer multilayer perceptron (MLP) projection head (output dimensions are 128). We set τ to 0.2 in all experiments. For data augmentations, we adopt SimCLR's basic augmentations for CIFAR-100. For Tiny ImageNet, we use MoCo v2's augmentation (to add gaussian blur to basic augmentations). The memory bank size is 4096. The momentum for the exponential moving average (EMA) update is 0.999. On CIFAR-100, models are trained for up to 300 epochs with a batch size of 128. On Tiny ImageNet, models are trained for 200 epochs with a batch size of 256. For optimization, we use stochastic gradient descent with a momentum of 0.9 and a weight decay of 1e-4 in all experiments. An initial learning rate is 0.125. For the learning rate scheduling, we use cosine decay schedule. For evaluation, we train a linear classifier for 90 epochs with a batch size of 128 in all experiments. An initial learning rate for linear evaluation is chosen among {0.5, 1.5, 2.5, 5, 15, 25, 35} and is dropped by a factor of 10 at 60%, 75%, and 90% of the training progress. We conducted the training on four Nvidia V100 GPUs.

3. **BYOL.** We use ResNet-18*. These networks are followed by the two-layer multilayer perceptron (MLP) projection head (output dimensions are 128). The momentum for the exponential moving average (EMA) update is 0.999. We do not symmetrize the BYOL loss. For data augmentations, we adopt SimCLR's basic augmentations. Models are trained for up to 300 epochs with a batch size of 128. For optimization, we use stochastic gradient descent with a momentum of 0.9 and a weight decay of 1e-4 in all experiments. An initial learning rate is 0.125. For the learning rate scheduling, we use cosine decay schedule. For evaluation, we train a linear classifier for 90 epochs with a batch size of 128 in all experiments. An initial learning rate for linear evaluation is chosen among {0.5, 1.5, 2.5, 5, 15, 25, 35} and is dropped by a factor of 10 at 60%, 75%, and 90% of the training progress. We conducted the training on four Nvidia V100 GPUs.

The setup of PDA, NDA and PNDA for SimCLR, MoCo v2 and BYOL are the same as above, respectively.

References

[1] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, 2020.

[2] Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. Accurate, large minibatch sgd: Training imagenet in 1 hour. *arXiv preprint arXiv:1706.02677*, 2017.

[3] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. In *ICLR*, 2017.

[4] Yang You, Igor Gitman, and Boris Ginsburg. Large batch training of convolutional networks. *arXiv preprint arXiv:1708.03888*, 2017.