Appendix for "AdCo: Adversarial Contrast for Efficient Learning of Unsupervised Representations from Self-Trained Negative Adversaries"

In this appendix, we further analyze the impact of several factors on the performance of the learned representation. We will demonstrate that

- Symmetrizing the contrastive loss has been proved effective in improving the performance of downstream tasks for existing models such as SimCLR [7] and BYOL [17]. For a fair comparison, we also adopt it in AdCo, and show that its top-1 accuracy can be increased by $1.5 \sim 2.0\%$ that achieves the state-of-the-art result compared to its asymmetric counterpart.

- AdCo is insensitive to the numbers of negative adversaries to pre-train the model. By reducing the number of negative samples from 65,536 to 8,192 (an eighth of the former size), AdCo has an unaffected top-1 accuracy. It shows that AdCo does not depend on a large amount of negative samples to pre-train the network.

- We also attempt to answer an emerging question of if we still need the contrastive learning and the associated negative examples to pretrain a deep network while the BYOL [17] does not rely on them any more. Specifically, the BYOL needs an extra MLP predictor with the same number of parameters as the negative examples that can be directly trained end-to-end in the AdCo. After comparing the computing costs among different methods, we show that the AdCo and its negative samples can be trained with almost 20% less GPU time with a much smaller batch size than the BYOL and its MLP predictor.

A. Symmetric Loss

Symmetrizing the loss in self-supervised learning has demonstrated improved performances for various models. For example, in SimCLR, a pair of positive samples augmented from the same image can exchange their roles in the contrastive loss by viewing one as a query and the other as a key and vice versa. Similarly, BYOL can also swap the inputs to its two branches, giving rise to a symmetric MSE term to minimize. In both SimCLR and BYOL, the symmetric loss has successfully improved the performance of the learned representation in downstream tasks.

We show in Table 5 that the performance of AdCo can also be improved by $1.5 \sim 2.0\%$ by symmetrizing its contrastive loss to update the representation network with the other factors (e.g., the number of negative samples and batch size) fixed. Thus, we will apply the symmetric loss in the following study when comparing it with the other models.

For a fair comparison, all results in Table 5 are obtained based on single-crop augmentations. We also tested the symmetric loss with multi-crop augmentations over 200 epochs of pre-training, but found that the symmetric loss only marginally improved the top-1 accuracy from 73.2% to 73.6%. We hypothesize that multi-crop augmentations have already leveraged multiple pairs of positive examples for the same image, and the benefit would vanish by symmetrizing the loss over those multi-crop augmentations.

B. Numbers of Negative Adversaries

One of the most important factors that could impact the performance of the learned representation on downstream tasks is the number of negative samples. For the sake of a fair comparison with the other SOTA contrastive models particularly MoCo v2, we have fixed it to $65,536$ in experiments. Here we will study if and how a smaller number of negative samples will impact the model performance.

The results in Table 5 show that the top-1 accuracy of AdCo is almost unaffected when the number of negative samples is reduced from 65,536 to merely 8,192. This suggests that the AdCo is insensitive to the change in the size of negative samples, no matter whether the symmetric loss is applied. Indeed, the top-1 accuracy of AdCo only varies by $0.2 \sim 0.4$ as the negative samples decrease to an eighth of the original size.

We hypothesize that this is attributed to the efficient adversarial training that results in more informative negative samples to self-supervise the network in AdCo. In other words, even a smaller number of negative adversaries are sufficiently representative to cover the learned representation in the embedding space, and there is no need to bring in too many negative samples to learn a high-performing contrastive model.

C. Do We Still Need Negative Samples?

MoCo v2 and BYOL are two state-of-the-art self-supervised models in literature. In Table 5, we compare the AdCo with them in terms of both accuracy and efficiency. The results show that both AdCo and BYOL outperform the compared models when the symmetric loss is applied.

Here, an insightful question may emerge – while the BYOL removes the need of negative samples in self-training a deep network, do we still rely on the contrastive learning and the associated negative samples for network pretraining?

When we proceed to further compare BYOL and AdCo, we note that although BYOL does not explicitly use any negative samples, it depends on an extra prediction MLP

Table 5: Top-1 accuracy under the linear evaluation on ImageNet with the ResNet-50 backbone. The table compares the methods over 200 epochs of pretraining with various batch sizes and numbers of negative samples. We also evaluate the impact of symmetric loss on these methods. The results show that with a smaller batch size, AdCo achieves the same top-1 accuracy to BYOL, while the latter needs almost 20% more GPU time to pretrain the model. AdCo also has a stable top-1 accuracy under different numbers of negative samples. All results are reported with a single-crop augmentation for a fair comparison.

Method	Symmetric Loss	Batch Size	#Neg. Samples	Top-1 Acc.	(GPU · Time) /epoch
SimCLR [7]	✓	8192	-	67.0	1.92
MoCo v2 [8]		256	65536	67.5	2.12
BYOL [17]	✓	4096	-	70.6	4.10
SimSiam [9]	✓	256	-	70.0	-*
AdCo		256	65536	68.6	2.26
AdCo		256	16348	68.6	2.24
AdCo		256	8192	68.4	2.24
AdCo	✓	256	65536	70.6	3.50
AdCo	✓	256	16384	70.2	3.46
AdCo	✓	256	8192	70.2	3.45

*Although no GPU time was reported on SimSiam, it should be on par with BYOL as its variant due to the similar network architecture and pre-training process.

to predict the embedding of an augmented view. There are empirical evidences [9] showing that such a MLP predictor plays an indispensable role in obtaining competitive results in experiments. Similarly, AdCo also has an extra network component by treating the trainable negative samples as an additional single neural layer attached onto the network to be pre-trained. In this sense, for a fair comparison, both BYOL and AdCo contain some extra model parameters. With the help of AdCo, it is now possible to directly train these parameters associated with negative samples end-to-end. While the number of parameters in the MLP predictor of BYOL is about 1M, the AdCo contains the same amount of trainable parameters associated with $8,192$ negative examples.

However, by comparison, BYOL needs almost 20% more GPU time than AdCo to pre-train the network. BYOL also relies on a larger batch size of $4,096$ than AdCo that has a much smaller batch size of 256 to achieve a competitive top-1 accuracy. This shows that AdCo can be trained in a more efficient fashion than BYOL.