Decoupled Adversarial Contrastive Learning for Self-supervised Adversarial Robustness

Chaoning Zhang*[1], Kang Zhang*[1], Chenshuang Zhang[1], Axi Niu[2], Jiu Feng[3], Chang D. Yoo[1], and In So Kweon[1]

[1] Korea Advanced Institute of Science and Technology (KAIST), Daejeon, Korea
chaoningzhang1990@gmail.com, zhangkang@kaist.ac.kr
[2] Northwestern Polytechnical University, Xi'an, China
[3] Sichuan University, Chengdu, China

Details of SLF and AFF. We follow the setting in AdvCL [2]. For standard linear fine-tuning (SLF), a linear classifier is applied on the top of a fixed encoder f_θ. This linear layer is trained with SGD for 25 epochs with batch size 512, initial learning rate 0.1, momentum 0.9, and weight decay 2e-4. The learning rate decays by 10 times at epoch 15 and 20. For the adversarial full fine-tuning (AFF), we adopt adopt the SOTA TRADES loss [5]. During training, to generate adversarial examples, we use 10-step ℓ_∞ PGD attack with $\epsilon = 8/255$ and the whole encoder f_θ with linear classifier will be trained and updated for 25 epochs. The learning rate schedule is the same as that of SLF.

Setup for DeACL at first stage. For the SSL at the first stage of our DeACL, we follow the practices in SoloLearn [1] to train the SimCLR encoder. We adopt SGD optimizer with momentum 0.9 and weight decay 1e-5. We train the model for 1000 epochs with a batch size of 256 on a single GPU. In the first 10 epochs, we use a linear warmup learning rate then decay learning rate following cosine decay schedule without restarts [3]. The projector consists of two linear layers with a ReLU activation function between them. The adopted augmentations include random resized crop, random color jittering, random grayscale conversion, random horizontal flip, etc. For more details, please refer to [1].

Table 1: Ablation study on the loss function.

Case	SLF		
	AA	RA	SA
Trades+Cossim (DC-SSL)	45.31	53.95	80.17
Trades+KL	10.55	10.00	10.00
Madry+Cossim	41.50	50.79	82.68

Ablation study on the loss of function. In supervised adversarial training, Mardy-AT [4] and Trades-AT [5] are the two most commonly used frameworks with different loss functions. Our loss function is similar to that in Trades-AT but replaces the KL divergence distance with the cosine similarity one. As

[*] Equal Contribution.

summarized in Table 1, such a distance replacement is beneficial for performance. Moreover, the loss in Madry-AT, where only AEs are used to train the encoder, is not as competitive as that in Trade-AT.

Table 2: Ablation study on the trade-off factor.

λ	SLF		
	AA	RA	SA
1	41.50	51.63	81.81
2	45.31	53.95	80.17
3	45.61	54.51	78.11
5	44.82	55.10	75.70

Ablation study for the trade-off factor. Table 2 reports the influence of λ of Eq 4 in the main manuscript. It shows that a higher λ tends to increase RA but decrease SA, which is well expected. Overall, taking the trade-off between robustness and accuracy into account, we set λ to 2.

Influence of initializing the student model with pretrained weights. We investigate the influence of initializing the student model with pretrained weights for our DeACL at the second stage. The weights refer to those pretrained at the first stage of our DeACL. As shown in Figure 1, loading the pretrained weights into the student model makes the convergence faster in the beginning, which is well expected. It also yields a small performance boost in the end.

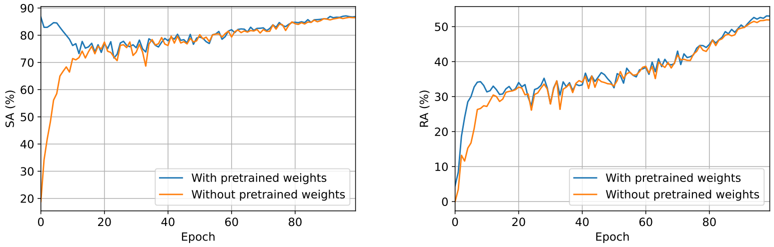

Fig. 1: Influence of initializing the student model with pretrained weights. SA (left) and RA (right) trend during the second stage of DeACL.

References

1. da Costa, V.G.T., Fini, E., Nabi, M., Sebe, N., Ricci, E.: Solo-learn: A library of self-supervised methods for visual representation learning. JMLR (2022)
2. Fan, L., Liu, S., Chen, P.Y., Zhang, G., Gan, C.: When does contrastive learning preserve adversarial robustness from pretraining to finetuning? NeurIPS (2021)
3. Loshchilov, I., Hutter, F.: Sgdr: Stochastic gradient descent with warm restarts. arXiv preprint arXiv:1608.03983 (2016)
4. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., Vladu, A.: Towards deep learning models resistant to adversarial attacks. In: ICLR (2018)
5. Zhang, H., Yu, Y., Jiao, J., Xing, E.P., Ghaoui, L.E., Jordan, M.I.: Theoretically principled trade-off between robustness and accuracy. In: ICML (2019)