Rethinking the Effect of Data Augmentation in Adversarial Contrastive Learning

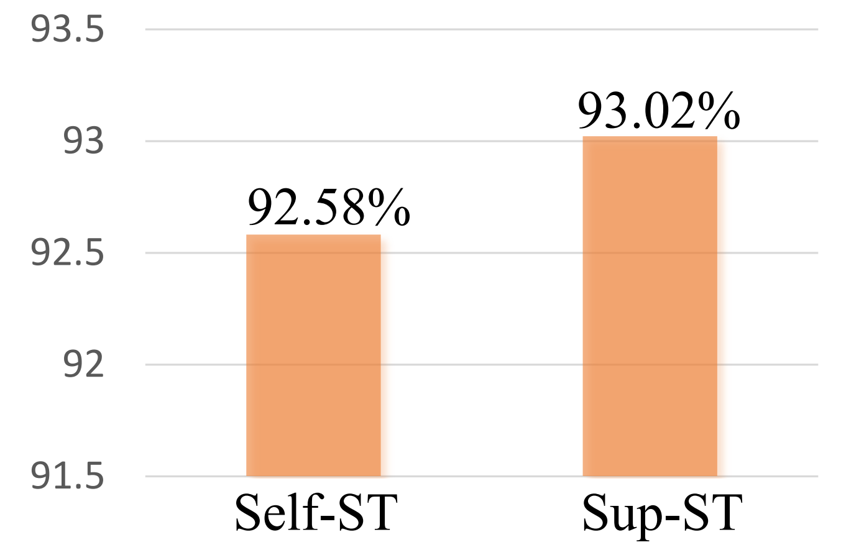
Rundong Luo*, Yifei Wang*, Yisen Wang†

Peking University

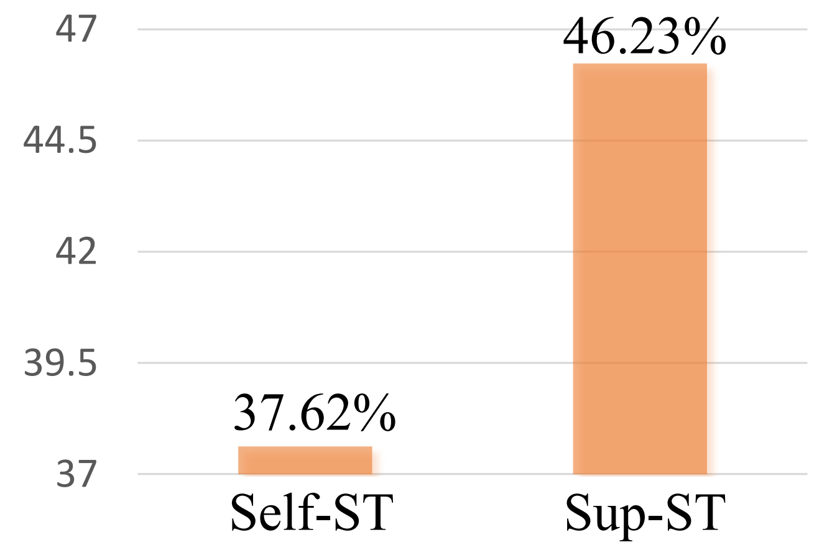
Motivation

Huge performance gap between supervised and self-supervised adversarial training

- Performance gap between self-supervised and supervised standard training (self/sup-ST): less than 1%. [1, 2]
- Performance gap between self-supervised and supervised adversarial training (self/sup-AT): more than 8%. [3, 4]



Standard Training



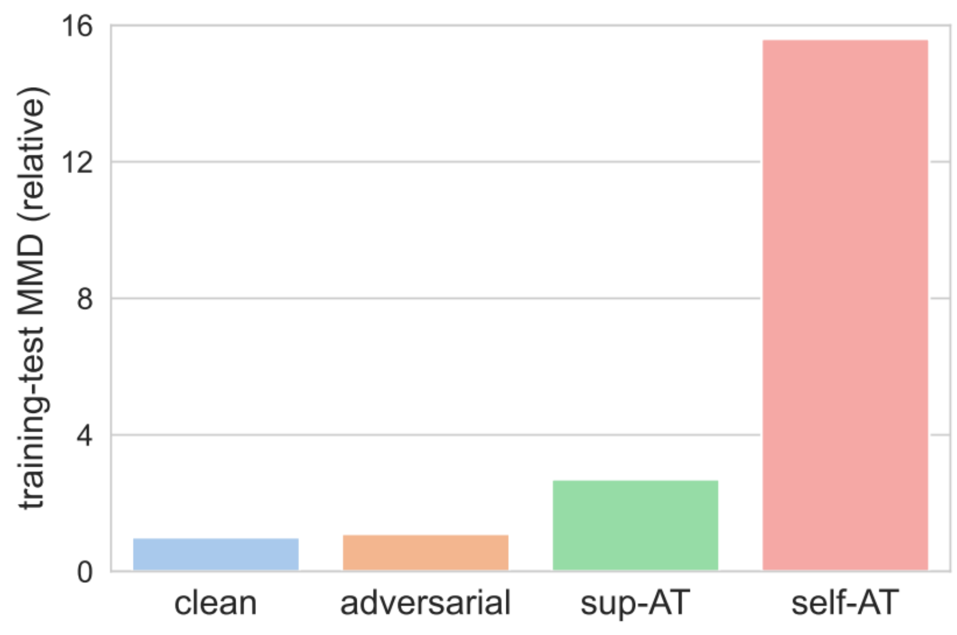
Adversarial Training

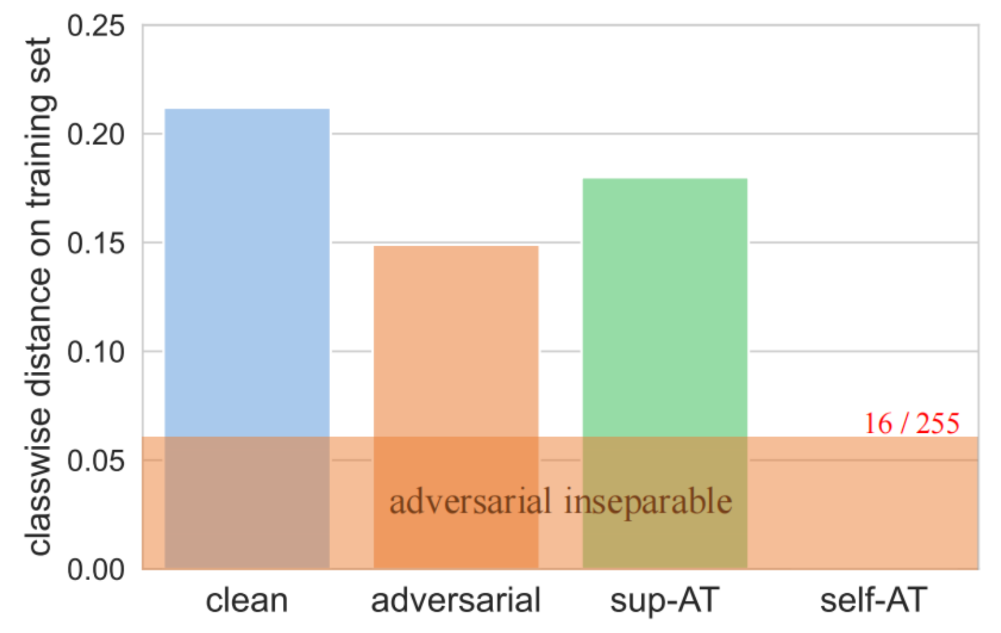
What is the key factor that prevents self-AT from obtaining comparable robustness to sup-AT?

Contributions

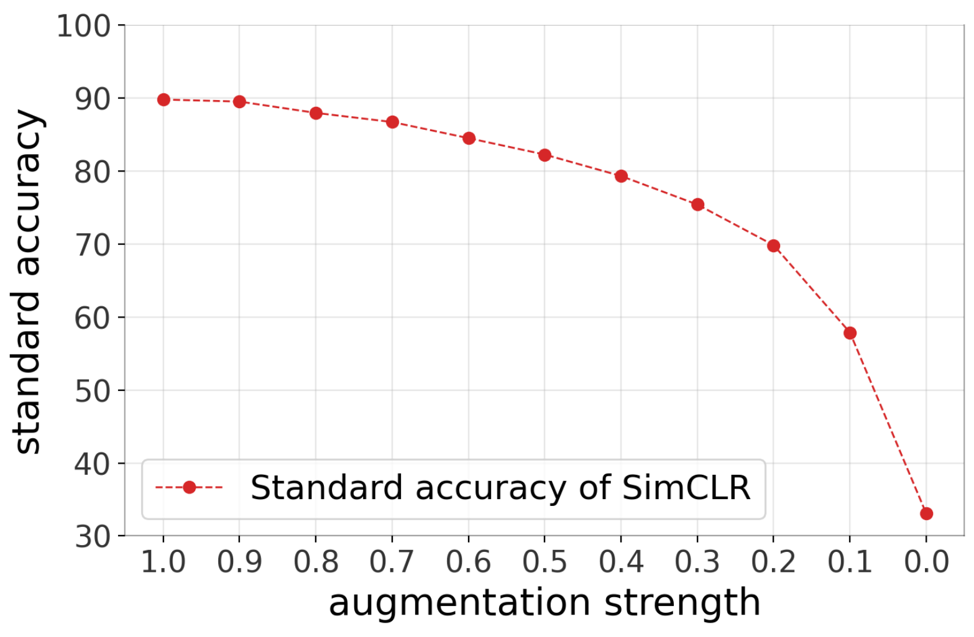
- **Analysis: Fierce data augmentation hinders self-AT.** We reveal the reason behind the robustness gap between self-AT and sup-AT, *i.e.*, the widely adopted aggressive data augmentations in self-supervised learning may bring the issues of training-test distribution shift and class inseparability.
- **Method: Dynamic Augmentation Strategy.** We propose a dynamic augmentation strategy along the training process to balance the need for strong augmentations for representation and mild augmentations for robustness, called Dynamic Adversarial Contrastive Learning (DYNACL) with its variant DYNACL++.
- **State-of-the-art Results.** Experiments show that our proposed methods improve clean accuracy and robustness over existing self-AT methods by a large margin. Notably, DYNACL++ enhances the robustness of ACL [3] from 37.62% to 46.46% on CIFAR-10, which is even slightly better than vanilla sup-AT.

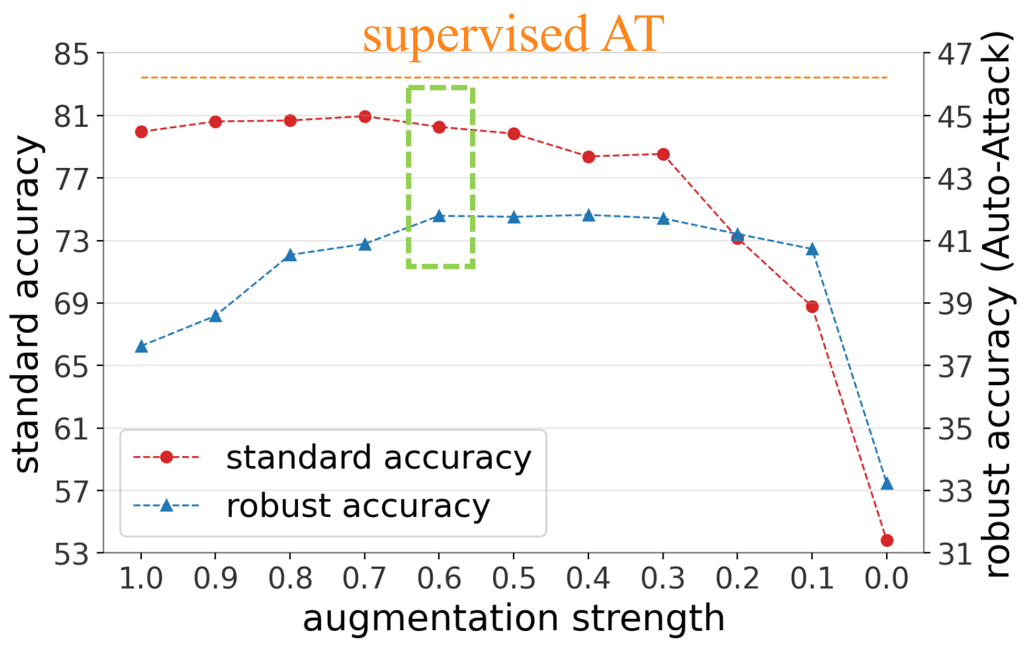
Discoveries





- Strong augmentation in self-AT enlarges the training-test distribution gap and the classwise distance, thus hurting the model's robustness.





- Decreasing the augmentation strength in contrastive learning hurts the performance.
- Strong augmentation is harmful to self-ST, and setting a medium aug. strength could bring improvements. However, further decreasing the augmentation strength is detrimental.

Dilemma in self-AT: Neither strong nor weak augmentation strength is suitable enough.

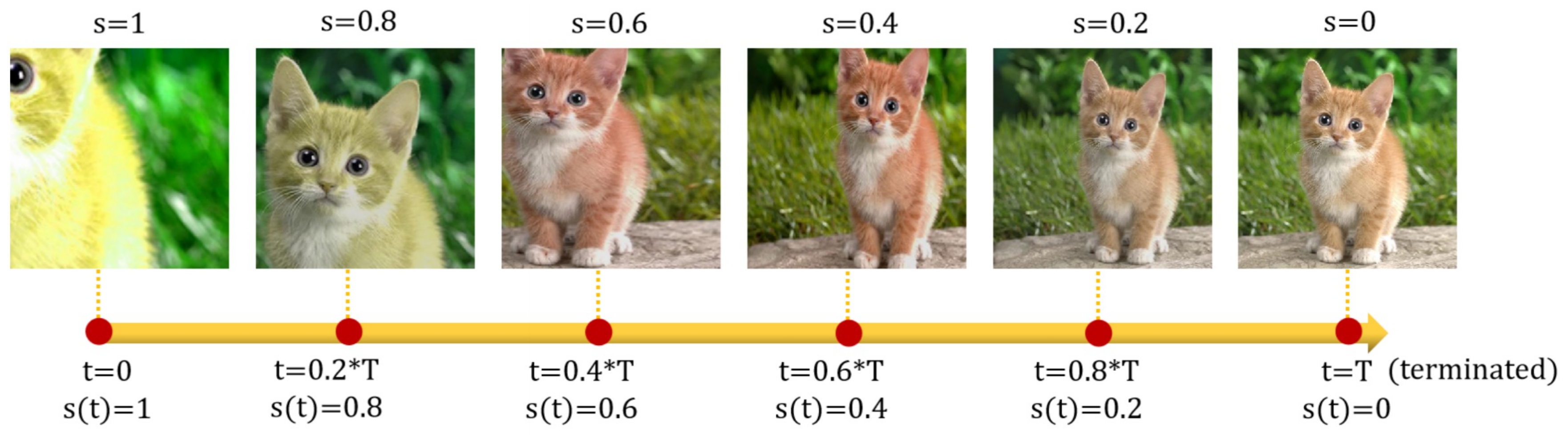
Method

DYNACL: Dynamic Adversarial Contrastive Learning

- We propose to treat the augmentation strength s as a dynamic parameter $s(t)$ that varies along current epoch t.
- Specifically, we gradually anneal s from 1 to 0 throughout the training process.
- This way, the model could learn meaningful representations at earlier training stages. Afterward, aug. annealing mitigates the distribution gap and narrow classwise distance, thus benefiting robustness.

$$s(t) = 1 - \lfloor \frac{t}{K} \rfloor \cdot \frac{K}{T}, \quad t = 0, \ldots, T-1 \tag{1}$$

T: total training epochs, K: data reload frequency, t: current epoch



For the t-th epoch, our loss is formulate as:

$$\mathcal{L}_{\text{DYNACL}}(g; t) = (1 - w(t))\mathcal{L}_{\text{NCE}}(g_c; s(t)) + (1 + w(t))\mathcal{L}_{\text{AdvNCE}}(g_a; s(t)).$$

g_a, g_c: backbone routes, $w(t)$: dynamic weight (2)

DYNACL++: DYNACL with fast post-processing

- Clustering

$$\{\hat{y}_i\} = \text{k\_means}(\{x_i\}; g_c) \tag{3}$$

- Pseudo Adversarial Training

$$\mathcal{L}_{PAT}(h \circ g_a) = \mathbb{E}_{\bar{x},y}\left\{\ell_{CE}(f(\bar{x}), y) + \max_{\delta_{\bar{x}} \in \Delta} KL(f(\bar{x}) \| f(\bar{x} + \delta_{\bar{x}}))\right\} \tag{4}$$

Experiments

- Comparative results under the self-supervised standard linear finetuning protocol and semi-supervised learning.

Pretraining Method	CIFAR-10		CIFAR-100		STL-10	
	AA(%)	SA(%)	AA(%)	SA(%)	AA(%)	SA(%)
Sup-AT	46.23	84.35	23.27	58.98	29.21	49.38
RoCL	26.12	77.90	8.72	42.93	26.51	**78.19**
ACL	37.62	79.32	15.68	45.34	33.24	71.21
AdvCL	37.46	73.23	15.45	37.58	45.26	72.11
DYNACL (ours)	45.04	77.41	19.25	45.73	46.59	69.67
DYNACL++ (ours)	**46.46**	**79.81**	**20.05**	**52.26**	**47.21**	70.93
AdvCL (+ImageNet)	42.57	80.85	19.78	48.34	N/A	N/A

Label Ratio	UAT++			ACL			**DYNACL++ (ours)**		
	AA(%)	RA(%)	SA(%)	AA(%)	RA(%)	SA(%)	AA(%)	RA(%)	SA(%)
1% labels	N/A	30.46	41.88	45.65	50.46	74.76	**46.95**	**51.30**	**76.77**
10% labels	N/A	50.43	70.79	45.47	50.01	75.14	**48.56**	**53.00**	**78.34**

DYNACL surpasses existing self-AT baselines and achieves better results than its vanilla sup-AT [4] counterpart.

Robustness is measured by Auto-Attack (AA) [5]. SA stands for standard accuracy.

References

[1] Victor Guilherme Turrisi da Costa, Enrico Fini, Moin Nabi, Nicu Sebe, and Elisa Ricci. solo-learn: A library of self-supervised methods for visual representation learning. JMLR, 2022.
[2] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In ICML, 2020.
[3] Ziyu Jiang, Tianlong Chen, Ting Chen, and Zhangyang Wang. Robust pre-training by adversarial contrastive learning. In NeurIPS, 2020.
[4] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. arXiv, 2017.
[5] Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In ICML, 2020.