Soften to Defend: Towards Adversarial Robustness via Self-Guided Label Refinement

Zhuorong Li[*1] **Daiwei Yu**[*1] Lina Wei[1] Canghong Jin[1] Yun Zhang[1]
Sixian Chan[2]

[1]Hangzhou City University, Hangzhou, China

[2]Zhejiang University of Technology, Hangzhou, China

CVPR 2024

Outline

Severe overfitting in Adversarial Training

- Adversarial training (AT) has been viewed as the main stream learning paradigm for obtaining robust classifier, which minimizes the worst-case loss within an ϵ-neighborhood of the input space.

$$\min_{\theta} \mathcal{L}_{adv}(\theta) := \frac{1}{n} \sum_{i=1}^{n} \max_{\delta_i \in \Delta} \ell(x_i + \delta_i, y_i; \theta).$$

- AT suffers from robust overfitting (RO), characterized by a significant generalization gap in robust accuracy between the training and testing curves.

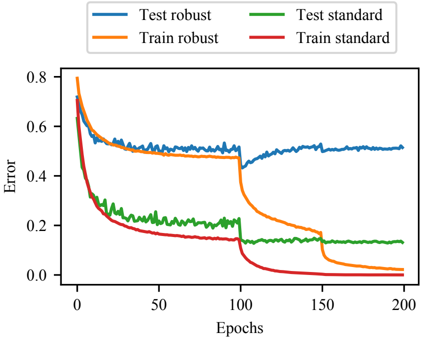


Figure: Robust overfitting

Previous theoretical understandings of RO

Theorem (Label noise in AT)

Assume $f(x)_y$ is $L-locally$ Lipschitz around x with Hessian bounded below, i.e., $\sigma_{min} \leq \sigma \leq \sigma_{max}$ and $\sigma_{min} = \inf_{z \in \mathcal{B}_\epsilon(z)} \sigma_{min}(\nabla^2 f(z)_y) > 0$. With probability $1-\delta$, we have

$$p_e(\mathcal{D}') \geq \frac{\epsilon}{2}(1 - q(\mathcal{D}))\frac{\sigma_{min}}{L} - \frac{\epsilon}{4}\sigma_{max} - \sqrt{\frac{1}{2N}\log\frac{2}{\delta}}$$

where σ^2 is the smallest eigenvalue of $\boldsymbol{\kappa}$.

- Assigned labels of adversarial examples are simply **inherited** from their clean counterparts.
- It suggests that as long as a training set is augmented by adversarial perturbation, but with assigned labels unchanged, **label noise emerges**.
- To reduce the label distribution mismatch, [2] rectify model probability with an adversarially trained teacher, which has **exacerbated** the consumption of computing resources.

Understanding RO through the lens of Noisy Label Learning

Does there exist more hands-off and hassle-free mitigation for robust overfitting?

Since the training process teeming with label noise, we could **directly** take *noisy label learning* into account during adversarial training.

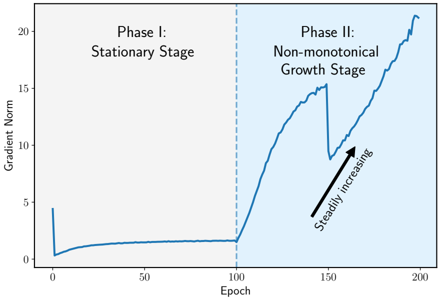
Lemma

Under PAC-Bayes framework, the expected cross entropy loss could be reformulated as follows:

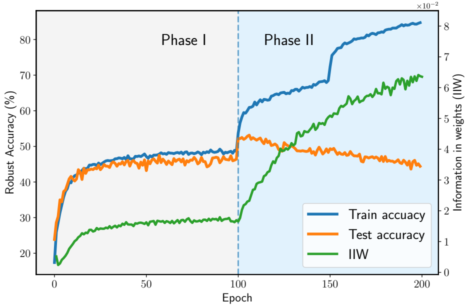
$$\begin{aligned}\mathcal{H}_f(\hat{y}|x,w) &= \mathbb{E}_{\mathcal{S}}\mathbb{E}_{w\sim Q(w|\mathcal{S})}\sum_{i=1}^{m}[-\log f(\hat{y}_i|x_i,w)] \\ &= \mathcal{H}(y|x) + \mathbb{E}_{x,w\sim Q(w|\mathcal{S})}\,\mathrm{KL}[p(y|x)\,\|\,f(\hat{y}|x,w)] - I(w;y|x)\end{aligned}$$

- Noisy labels viewed as the outlier of true label distribution can provide a **positive value of** $I(w;y|x)$ (**Information in weights**) as training goes.

Empirical perspective



(a) Gradient norm magnitude



(b) Generalization gap

- According to the LR decays, the training process could be divided into two stages: **(i) Stationary Stage (ii) Non-monotonical Growth Stage**.
- The abrupt increment of gradient norm, failing to converge to a constant, could be seen as an indicator of **memorization effects** (on **noisy labels**) during learning.
- Simultaneously, the behavior of the IIW exhibits **trends similar** to that of the gradient norm, which could be viewed as a characteristic of RO.

Could we reduce the IIW so as to mitigate RO?

Theorem

Let u be the uniform random variable with p.d.f $p(u)$. By using the composition in Lemma 1., there exists an interpolation ration λ between the clean label distribution and uniform distribution, such that

$$I(y^*; w|x') \lesssim I(y; w|x')$$

where $p(y^*|x', w) = \lambda \cdot p(y|x', w) + (1-\lambda) \cdot p(u)$ and the symbol $\lesssim$ means that the corresponding inequality up to an c-independent constant.

- For some type of soft label, there exists an excellent label distribution interpolation between clean label distribution and well-designed label distribution that could effectively reduce the IIW.

Method: Self-Guided Label Refinement

From Theorem 2, we note that some type of soft label can **reduce IIW**, thus **mitigating RO**. So we could rectify model prediction probability with **reliable** knowledge learned by **model itself**.

$$\begin{aligned} \mathbf{y} &= r \cdot \widetilde{p}_t + (1 - r) \cdot \mathbf{y}_{hard} \\ \widetilde{p}_t &= \alpha \cdot \widetilde{p}_{t-1} + (1 - \alpha) \cdot \widetilde{f}(x, x'; w_t) \end{aligned} \tag{1}$$

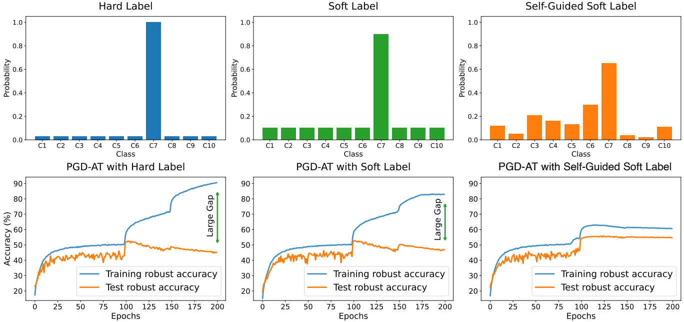


Figure: Robust accuracy of models employing different label assignment methods.

Results on CIFAR-10

Table: Test accuracy (%) of the proposed method and other methods on CIFAR-10 under the ℓ_∞ norm with $\epsilon = 8/255$ based on the ResNet-18 architecture.

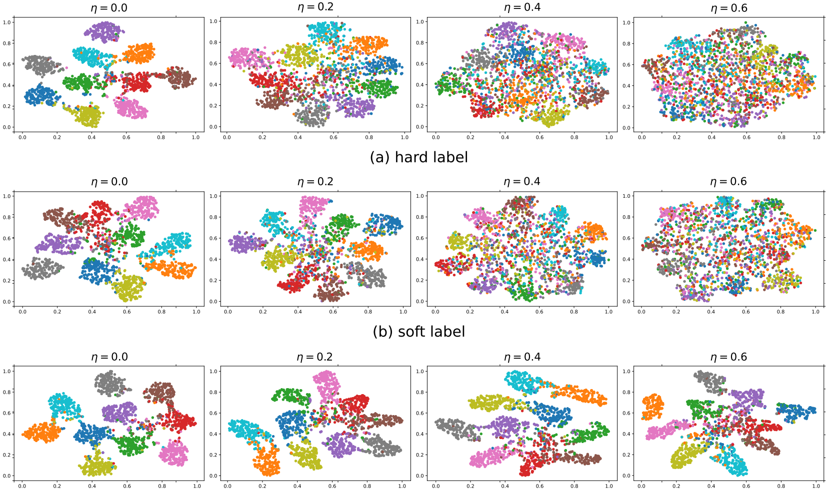
Method	Natural Accuracy			PGD-20			AutoAttack		
	Best	Final	Diff ↓	Best	Final	Diff ↓	Best	Final	Diff ↓
PGD-AT	80.7	82.4	-1.6	50.7	41.4	9.3	47.7	40.2	7.5
PGD-AT+LS	82.2	84.3	-2.1	53.7	48.9	4.8	48.4	44.6	3.9
PGD-AT+TE	82.4	82.8	-0.4	55.8	54.8	1.0	50.6	49.6	1.0
PGD-AT+SGLR	82.9	83.0	**-0.1**	**56.4**	**55.9**	**0.5**	**51.2**	**50.2**	1.0
AWP	82.1	81.1	1.0	55.4	54.8	0.6	50.6	49.9	0.7
KD-AT	82.9	**85.5**	-2.6	54.6	53.2	1.4	49.1	48.8	0.3
KD-SWA	**84.7**	85.4	-0.8	54.9	53.8	1.1	49.3	49.4	**-0.1**
PGD-AT + SGLR	82.9	83.0	**-0.1**	**56.4**	**55.9**	**0.5**	**51.2**	**50.2**	1.0

Results on other datasets.

Table: Clean accuracy and robust accuracy (%) of ResNet 18 trained on different benchmark datasets. All threat models are under ℓ_∞ norm with $\epsilon = 8/255$. The bold indicates the improved performance achieved by the proposed method.

Dataset	Method	Natrural Accuracy			PGD-20			AutoAttack		
		Best	Final	Diff ↓	Best	Final	Diff ↓	Best	Final	Diff ↓
CIFAR-10	AT	80.7	82.4	-1.6	50.7	41.4	9.3	47.7	40.2	7.5
	+SGLR	**82.9**	**83.0**	**0.1**	**56.4**	**55.9**	**0.5**	**51.2**	**50.2**	**1.0**
	TRADES	81.2	82.5	-1.3	53.3	50.3	3.0	49.0	46.8	2.2
	+SGLR	**82.2**	**83.3**	**-0.9**	**55.8**	**55.4**	**0.4**	**50.7**	**50.1**	**0.6**
CIFAR-100	AT	53.9	53.6	0.3	27.3	19.8	7.5	22.7	18.1	4.6
	+SGLR	**56.9**	**56.6**	**0.3**	**34.5**	**34.3**	**0.2**	**27.5**	**26.7**	**0.8**
	TRADES	**57.9**	56.3	1.7	29.9	27.7	2.2	24.6	23.4	1.2
	+SGLR	57.1	**57.4**	**-0.3**	**33.9**	**33.2**	**0.7**	**27.1**	**26.4**	**0.7**

Separable features of T-SNE plot under different noise rate.



(a) hard label

(b) soft label

(c) self-guided soft label

Conclusion

- We provide empirical and theoretical understanding on robust overfitting through the perspective of noisy label learning.
- We propose Self-Guided Label Refinement to obtain an informative label distribution, which achieves significantly improved clean and robust accuracy.

ArXiv: `https://arxiv.org/abs/2403.09101`