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 [2]Zhejiang University of Technology
{lizr, weiln, jinch, yunzhang}@hzcu.edu.cn, ydw.ccm@gmail.com, sxchan@zjut.edu.cn

Abstract

Adversarial training (AT) is currently one of the most effective ways to obtain the robustness of deep neural networks against adversarial attacks. However, most AT methods suffer from robust overfitting, i.e., a significant generalization gap in adversarial robustness between the training and testing curves. In this paper, we first identify a connection between robust overfitting and the excessive memorization of noisy labels in AT from a view of gradient norm. As such label noise is mainly caused by a distribution mismatch and improper label assignments, we are motivated to propose a label refinement approach for AT. Specifically, our Self-Guided Label Refinement first self-refines a more accurate and informative label distribution from over-confident hard labels, and then it calibrates the training by dynamically incorporating knowledge from self-distilled models into the current model and thus requiring no external teachers. Empirical results demonstrate that our method can simultaneously boost the standard accuracy and robust performance across multiple benchmark datasets, attack types, and architectures. In addition, we also provide a set of analyses from the perspectives of information theory to dive into our method and suggest the importance of soft labels for robust generalization.

1. Introduction

Recent studies have reported that deep neural networks (DNNs) are vulnerable to adversarial examples, *i.e.*, malicious inputs perturbed by an imperceptible noise to confuse the classifier prediction [14, 25]. This vulnerability raises serious security concerns and motivates a growing body of works on defense techniques [18, 24, 39]. Adversarial training (AT) is arguably the most promising way to harden classifiers against adversarial examples, which directly augments the training set with adversarial examples [3, 18]. It is formulated as a min-max problem [18] to find model parameters w that minimize the adversarial risk:

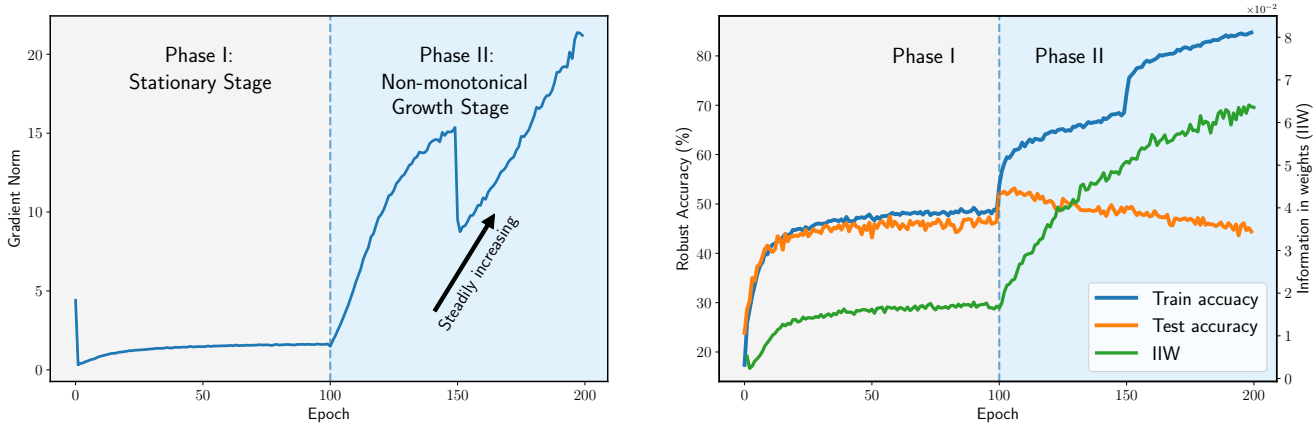
$$\min_{w} \mathcal{L}_{\mathcal{S}}(x', y; w) \tag{1}$$

where $\mathcal{L}_{\mathcal{S}}(x', y; w) = {}^1\!/_n \sum_{i=1}^{|\mathcal{S}|} \max_{x' \in \mathcal{B}_r(x)} \ell(f(x'_i; w), y_i)$ and $f(\cdot; w)$ is a model parameterized by w, $\ell(\cdot)$ is the loss function such as cross-entropy loss, and $\mathcal{B}_r(x)$ denotes the set of the allowed perturbations under the given metric space $M = (X, d)$ and the suitable radius $r > 0$, *i.e.*, $\mathcal{B}_r(x) = \{x + \delta \in \mathcal{X} : d(x, x + \delta) < r\}$.

However, most AT methods suffer from a dominant phenomenon that is referred to as "robust overfitting". That is, an adversarially trained model can reach almost 100% robust accuracy on the training set while the performance on the test set is much inferior, witnessing a significant gap of adversarial robustness [22]. Various regularization techniques including classic ℓ_1, ℓ_2 regularization and more advanced regularizations using data augmentation, such as Mixup [38] and Cutout [9], have been attempted to mitigate robust overfitting, whereas they are reported to perform no better than a simple early stopping [22]. However, early stopping raises another concern as the checkpoint of the best robustness and that of the best standard accuracy often do not coincide [5]. To outperform data augmentations and early stopping, regularizations specifically designed for robust training are thus proposed, to name a few, loss reweighting [27, 29, 37, 40] and weight smoothing [5, 33, 36, 37]. These regularization methods are likely to restrict the change of training loss by suppressing perturbations with respect to either the inputs x or the weights w, whereas few explorations attempt to combat robust overfittiing from the perspective of labels y.

Previous investigations on labels in AT [10, 35] emphasised the existence of label noise in AT to provide an understanding of robust generalization, for instance, from a classic bias-variance perspective. We take a step still further to investigate the memorization of label noise in AT, which is characterised by a non-monotonical increase of gradient norm, as illustrated in Fig. 1, resulting in robust overfitting. In light of our analyses, we are motivated to design a strategy

† Corresponding author.
* The first two authors contribute equally.



(a) Gradient norm magnitude (b) Generalization gap

Figure 1. In figure (a), we calculate the gradient norm of vanilla adversarially trained PreAct-ResNet 18 on CIFAR-10 for robustness against ℓ_∞ perturbations of radius $8/255$. In figure (b), we show the robust accuracy under PGD-20 attack under the same settings with figure (a). The gradient norm keeps non-monotonically ramping up when robust overfitting happens.

for label refinement to alleviate excessive memorization and thus the robust overfitting. To that end, we conduct an empirical experiment using different label assignments and analyse through the lens of learning curve, which is a useful indicator for the occurrence of robust overfitting. As shown in Fig. 2, the phenomenon of robust overfitting is presented, for instance, in the lower left of Fig. 2 for PGD-AT [18] with commonly used hard labels, and at a slightly reduced extent as shown in the lower middle of Fig. 2 with vanilla soft labels.

In this work, we propose a theoretical-grounded method to alleviate the memoization on over-confident labels during adversarial training and thus to combat robust overfitting. Our main idea is to resort to an alternative for label assignment. Particularly, motivated by the effectiveness of soft label in alleviating overfitting in standard training [26], our work generates more reliable labels automatically by incorporating predictive label distributions into the process of robust learning. It provides a promising way to inject distilled knowledge and to self-calibrate the adversarial training. Our method is conceptually simple yet significantly enhances the learning of deep models in adversarial scenarios. The key contributions are as follows:

- We first inspect the behaviour of deep models trained by AT when robust overfitting occurs. Specifically, we identify a connection between robust overfitting and the excessive memorization of noisy labels in AT through the lens of gradient norm (see Fig. 1). As such label noise is mainly due to unaccurate label distribution, we further investigate the effects of different label assingment methods on AT (see Fig. 2). These observations consistently implies a connection between robust overfitting and noisy hard labels. Exploring such connections could help to shed light on understanding and improving the robust learning of deep models.
- Upon the observation, we are motivated to propose a label refinement approach for AT. Specifically, our Self-Guided Label Refinement (SGLR) first self-refines accurate and informative label distribution from the over-confident hard labels, and then it calibrates the training by dynamically incorporating knowledge from self-distilled models into the current model, requiring no external teacher models nor modifications to the existing architecture.
- We verify the effectiveness of SGLR through extensive evaluations. Overall, experimental results show that the proposed method consistently improves the test accuracy over the state-of-the-art on various benchmark datasets against diverse adversaries. Moreover, our approach can achieve robust accuracy up to 56.4% and close the generalization gap to merely 0.4%, significantly mitigating the overfitting issue and thus being able to pushing up the adversarial robustness.

2. A closer look at robust overfitting

Robust overfitting has been prevalent across various datasets and models [22]. Then, crops of empirical and theoretical studies have emerged to analyze this phenomenon through the lens of *loss value* [37], *training data* [10] and *learned features* [30]. However, a comprehensive underlying mechanism of robust overfitting still remains an enigma. In this section, we revisit robust overfitting from a perspective of the established information theory and identify a "*memorization effect*" over the course of adversarial training, which is akin to standard training as oberved in [6, 34]. First and foremost, we conduct some observations when robust overfitting occurs via analysing the change of the gradient norm of adversarial loss with respect to model weight, *i.e.*, $\|\nabla_w L(x', y, w)\|_2$. As shown in Fig. 1 (a), we may note that the gradient norm holds nearly the constant and then keeps ramping up non-monotonically. This increasing behavior of gradient norm

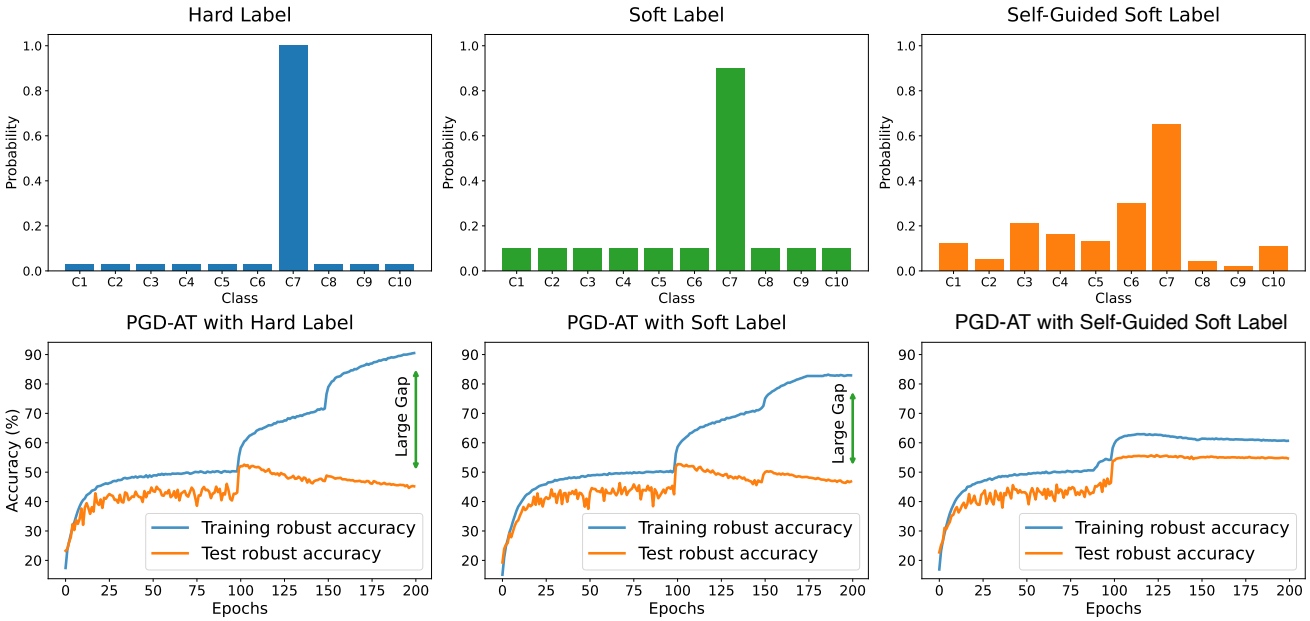


Figure 2. Robust accuracy of models employing different label assignment methods in adversarial training.

encountered with the learning rate (LR) decays. So we divide the training process into two phases according to the LR decays: (1) ***Stationary stage***; (2) ***Non-monotonical growth stage***.

In stationary stage, as illustrated in Fig. 1 (b), the learning curve (orange line) and test robust accuracy (blue line) ramp up at almost the same pace and maintain a very small generalization gap, *i.e.*, the divergence between the training and test accuracy, implying that the model keeps learning efficient robust features as the training progresses and accrodingly hovers a small constant gradient norm. However, a significantly expanding generalization gap can be witnessed after an inflection point, where the LR decays. We refer the period afterwards as "Non-monotonical growth stage", as a prominent characteristic of this stage is that the gradient norm keeps growing non-monotonically and does not converge to a constant, even after the training has ended.

It is worth noting that this trend of gradient norm greatly contradicts with that of the conventional ERM training, where zero gradient norm can be reached at the end of training. Nonetheless, such behaviour of gradient norm is in line with ERM when significant noisy labels surrounding, as observed in [11, 31]. Specifically, such growing trend is interpreted as an indicator that the training has entered a "memorization routine", where the model firstly fits on training samples with clean labels, then gradually over-fits samples with noisy labels. Later in this stage, the test accuracy on clean data will go down whereas the training accuracy keeps going up. Similarly, we are able to identify a non-monotonical growth in adversarial training, as depicted in Fig. 1 (a). Remarkably, this divergent gradient norm as the adversarial training progressing, exactly accompanies with the enlargement of generalization gap that is depited in Fig. 1 (b), which reveals the overfitting problem. Thus, we come to a conjecture that the memorization effect is also the curprit of the increasing gradient norm in the scenario of robust training and further induces robust overfitting. To that end, we first provide proof for our hypothesis that advesarial training memorizes samples with noisy labels mainly in the non-monotonical divergence phase.

Following [1], the expected cross entropy loss could be decomposed into three terms according to PAC-Bayesian framework:

$$\begin{aligned}
&\mathcal{H}_f(\hat{y}|x,w) \\
&= \mathbb{E}_{\mathcal{S}}\mathbb{E}_{w\sim Q(w|\mathcal{S})}\sum_{i=1}^{m}[-\log f(\hat{y}_i|x_i,w)] \\
&= \mathcal{H}(y|x) + \mathbb{E}_{x,w\sim Q(w|\mathcal{S})}\mathrm{KL}[p(y|x)\,\|\,f(\hat{y}|x,w)] - I(w;y|x)
\end{aligned} \tag{2}$$

In conventional ERM training, minimizing $-I(w;y|x)$ relies on the cross entropy loss between the prediction $f(\hat{y}|x,w)$ and the true label distribution $p(y|x)$. Noisy label viewed as the outlier of a true label distribution can provide a positive value of $I(w;y|x)$. This term essentially quantifies the extent to which label information overlaps the weights of the model. In other words, the presence of noisy labels will reduce the finite degree of freedom with respect to weights. The reason for this reduction is that model attempts to accommodate the noise labels in the training data, inducing a potential overfitting. Different from the conventional ERM training, adversarial training requires perturbed samples x', which are usually obtained by solving a multi-step maximization problem. Accordingly, we use the notation $p(y'|x')$ for the true distribution of adversarial samples. Based on the

Eq. (2), we have

$$\begin{aligned}\mathcal{H}_f(\hat{y}|x',w) =& \mathbb{E}_{\mathcal{S}}\mathbb{E}_{w\sim Q(w|\mathcal{S})}\sum_{i=1}^{m}\left[-\log f(\hat{y_i}|x'_i,w)\right] \\ =& \mathcal{H}(y'|x') - I(w;y'|x') \\ &+\mathbb{E}_{x',w\sim Q(w|\mathcal{S})}\mathrm{KL}[p(y'|x') \parallel f(\hat{y}|x',w)]\end{aligned} \tag{3}$$

As aforementioned, the adversarial perturbation cause a mismatch between the label distributions of the perturbed data and their origins, as $p(y'|x') \neq p(y|x)$. However, during adversarial training, $p(y'|x')$ is often simply replaced by $p(y|x')$, that is, the perturbed labels are directly inherited from their origins. This would inevitably exert influence on the label memorization and results in performance degradation.

We are interested in the memoization of noisy labels in adversarail training, which can be characterised by $I(w;y'|x')$ and refered to as "Information In Weights" (IIW). The training dynamics can also be captured through the lens of IIW. However, computing the value of $I(w;y'|x')$ is as difficult as fitting the model itself. Basing on the chain rule of mutual information, we could approach to our desired result via the upper bound, *i.e.*, $I(\mathcal{D};w) = I((x,y);w) = I(x;w) + I(y;w|x)$. By using the positivity property of mutual information, we have $I(y;w|x) \leq I(D;w)$. Then we could take an approximation under some Gaussian assumptions, as suggested by Wang et al. [32], to estimate $I(w;y'|x')$ and further depict the learning behaviour of model, which is illustrated by the green line in Fig. 1 (b). The increasing IIW supports our hypothesis that adversarially trained model mainly overfits the noisy labels in the non-monotonical growth stage. Upon the theoretical analysis, it is natural that we proposed to prevent models from excessively memorising noisy labels by means of IIW reductions. In this work, we propose to reduce IIW through soft labels, which is specifically effective in lowering mutual information and thus can be expected to weaken the memorization under the distribution mismatch.

Theorem 1 (Soft label could reduce the IIW) *Let u be the uniform random variable with p.d.f $p(u)$. By using the composition in Eq. (2), there exists an interpolation ration λ between the clean label distribution and uniform distribution, such that*

$$I(y^*;w|x') \lesssim I(y;w|x') \tag{4}$$

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.

The detailed proof could be found in Appendix A. Theorem 1 proves that there exists some kind of soft label that could reduces the information in weights. So the memorization effect caused by the label distribution mismatch could be effectively mitigated. Thus, to better suppress the memorization effect, we should provide more underlying information about the true label distribution than uniform distribution to facilitate a better interpolation of the soft label. In the following, we introduce our solution to estimate more accurate and informative soft labels for adversarial training.

3. Self-guided label refinement for adversarial training

3.1. Methodology

As discussed above, hard labels in adversarial training are uninformative but over-confident, and thus heavily impairs the generalization (see Fig. 2). To address this issue, we propose an alternative to one-hot labels for adversarial training. Specifically, our Self-Guided Label Refinement first utilizes the learned probability distributions to refine the over-confident hard labels, and then it guides the training process using the historical training model to obtain a calibrated prediction.

We begin by softening the over-confident hard labels. It is well acknowledged that label smoothing (LS) helps to calibrate the degree of confidence of a model and it is effective in improving the robustness in noise regimes [17]. The vanilla LS for a K-class classification problem can be formulated:

$$\mathbf{y} = \frac{r}{K}\cdot\mathbf{1} + (1-r)\cdot\mathbf{y}_{hard} \tag{5}$$

where, $\mathbf{y}_{hard}$ denotes labels encoded by a one-hot vector, the notation of $\mathbf{1}$ denotes all one vector, and $r \in [0,1]$ controls the smooth level. It is shown that LS serves as a regularizer as well as a confidence penalty [21] and therefore improves the generalization of the model in standard training. Unfortunately, when it comes to adversarail training, a direct combination with LS cannot guarantee reliable robustness, especially in the cases of strong perturbations[4, 20]. Specifically, as the uniform distribution is unlikely to match the underlying distribution, LS tends to introduce a bias that might hurt the robust generalization. Moreover, soft labels with identical probability over the false categories cannot reveal the semantic similarity.

To alleviate this artificial effect, we introduce our Self-Guided Label Refinement (SGLR) which utilizes the knowledge inferred by a trustworthy model itself to retains informative labels. The proposed SGLR can be formulated as:

$$\mathbf{y} = r\cdot f(x';w) + (1-r)\cdot\mathbf{y}_{hard} \tag{6}$$

where $f(x';w)$ is the logit output of the model parameterized by w on training data x'. To be noticed, the model f we referred to here is not pre-trained but rather on the training. Eq. (6) also in fact serves as a regularizer by integrating the knowledge extracted from the model with stastic information

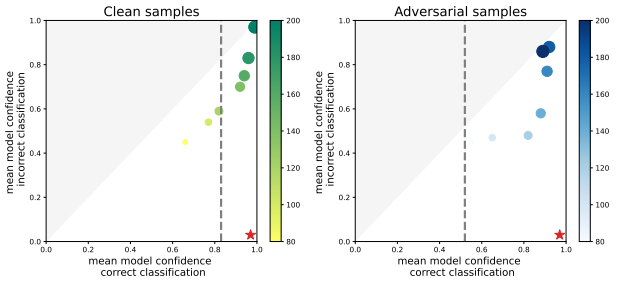


Figure 3. The mean confidence of model in the correct and incorrect predictions over clean and adversarial test sets.

that encoded by one-hot labels, thus, we suggest that this particular form of soft label has the potential to convey a better quantity of informative content.

Furthermore, according to Sanyal et al. [23], there exists a good interpolation of the noisy training dataset that could lead to a satisfactory generalization. Consequently, we aim to establish a valid interpolation between robust and non-robust features with the expectation that it strikes a balance between accuracy and robustness, as represented by the equation:

$$\widetilde{f}(x, x'; w) = \lambda \cdot f(x; w) + (1 - \lambda) \cdot f(x'; w) \tag{7}$$

Fig. 3 illustrates the major trajectory of PGD-AT training on CIFAR-10 dataset in terms of the confidence in both the correct and incorrect predictions, where the model dynamics are shown by a series of colored dots with darker colours and larger areas. Gray dashed lines mark the best standard / robust accuracy of the trained model, and the red star in the bottom right corner denotes an ideal model that with a high confidence in correct classification whereas a low one in the incorrect prediction. As training processes, model tends to assign increasing confidence to its correct predictions and also to those wrong predictions, as implied by the moving of the colored dots away from the dash line. In other words, the latter prediction of the model is not better calibrated and the model in the middle training stage could be of great help to reduce the expected calibration error. Therefore, we are motivated to utilize exponential moving average (EMA) that taking the moving average of history model prediction to obtain calibrated soft label. Intuitively, EMA considers recent proposals from the current state while retaining some influence from previous information. This straightforward method is easy to deploy and incurs negligible training overhead. Then the dynamic updating of soft label could be defined as follows:

$$\begin{aligned} \mathbf{y} &= r \cdot \widetilde{p}_t + (1 - r) \cdot \mathbf{y}_{hard} \\ \widetilde{p}_t &= \alpha \cdot \widetilde{p}_{t-1} + (1 - \alpha) \cdot \widetilde{f}(x, x'; w_t) \end{aligned} \tag{8}$$

3.2. Connection to symmetric cross entropy

Cross Entropy (CE) is the prevalent loss function for deep learning with remarkable success. However, as the noisy label is tipping the training process into overfitting, CE loss has taken a nasty blow and is prone to fitting noise. Inspired by the symmetric Kullback-Leibler (KL) Divergence, Wang et al. [28] proposed a more flexible loss, *i.e.*, symmetric cross entropy (SCE), to strike a balance between sufficient learning and robustness to noisy labels. Moreover, it reveals that predictive distribution exploited by the model is superior to one-hot label distribution for the most part. Both SCE loss and our method utilize the informative knowledge inferred by the model and Proposition 1 provide such a close link between them.

Proposition. 1 *Let ℓ_{sce} be the SCE loss function and γ represents the sum of CE and reverse CE loss. When $\gamma \to 1$, then our methods can also be written as:*

$$\ell_{sglr} = \ell_{sce} - \alpha \cdot \ell_{rkl}$$

where ℓ_{rkl} denotes the reverse KL divergence between labels and model predictions, i.e., $D_{KL}(p \,||\, q)$.

The proof of Proposition 1 can be found in Appendix A. Note that Proposition 1 indicates our method can be decomposed into two terms. The first term represents the SCE loss function when the weighted sum tends to 1. The second term indicates that our method rewards the distribution differences between predictions and labels. Though our study aims at mitigating robust overfitting, which is different from the scenario of SCE loss, there is no conflict between our method and SCE.

3.3. Comparison to knowledge distillation

Knowledge distillation (KD) [16] has been proven to be an effective way to compress models, which is able to remarkably beef up the performance of lightweight models. There is a huge scope for applying KD to adversarial training to improve the efficiency of AT. A crop of works [5, 13, 29, 41] have advanced exploring KD implicitly or explicitly in conjunction with AT. The efficacy of KD is attributed to the teacher model's informative knowledge, which guides the student to acquire a more similar knowledge representation and effectively capture inter-class relationships. By enforcing consistency in probability distributions, KD facilitates better learning outcomes for the student.

Proposition. 2 *Some KD methods, which minimize the distance of the feature map between the teacher and student model, belong to the family of our method. Let p_t be the prediction of the teacher model and then the KD could also be written as $\ell_{KD} = \mathbb{E}_{\widetilde{q}}\left[-\log p\right] = H(\widetilde{q}, p)$, where $\widetilde{q} = (1 - \alpha) \cdot q + \alpha \cdot p_t$.*

Proposition 2 shows that some KD revised the hard label via the knowledge p_t learned by the teacher. The proof of Proposition 2 is presented in Appendix A. Further, we view

Table 1. Comparison between our method and other KD methods.

	ARD [13]	RSLAD [41]	KD-SWA [5]	Ous
Trained teacher models	1	1	2	0
Standardly teacher independent	✓	✓	✗	✓
Adversarially teacher independent	✗	✗	✗	✓
Forward times in one training iteration	3	3	4	2

our method as special supervision, and smoothing the hard label could better reflect the similarities among classes. There are many differences between our method and other KDs in AT. The main differences are summarized in Tab. 1. Our method does not incur any extra computational cost as we do not involve teacher models, which is a boon throughout the whole training progress.

3.4. Tolerant to noisy label

In the following part, we'd like to delve into whether the proposed method is tolerant of noisy labels. We define the symmetric noise label as that true label y has probabilities $\eta_{x,\widetilde{y}} = p(\widetilde{y} \,|\, y, x)$ to flip into the wrong labels uniformly. The corresponding *noisy empirical risk* is:

$$R_{\mathcal{S}}^{\eta}(f) = \mathbb{E}_{\mathcal{S}}(1-\eta_x) \cdot \ell(f(x), y) + \sum_{i \neq y} \eta_{x,i} \ell(f(x), i)$$

where, η_x is the noise rate. We call a loss function noise-tolerant if and only if the global minimum f_{η}^{*} has the same probability of misclassification as that of f^* on the noise-free data.

Lemma. 1 *Given a symmetric loss function ℓ that it satisfies $\sum_{i=1}^{k} L(f(x), i) = C$ and C is some constant. Then ℓ is noise tolerant under symmetric label noise if noise rate η meets $\eta < 1 - \frac{1}{K}$.*

The loss condition theoretically guarantees the noise tolerance by risk minimization on a symmetric loss function following [12] and it shows that the global optimal classifier f^* on noise-free data remains the optimal even with the noisy label. Further, we can also derive the noise tolerance theoretically about our method from Theorem 2.

Theorem 2 *In a K-class classification problem, our method $\widetilde{\ell}$ is noise-tolerant under symmetric or uniform label noise if noise rate $\eta < 1 - \frac{1}{K}$. And if $R(f^*) = 0$, $\widetilde{l}$ is also noise-tolerant under asymmetric or class-dependent label noise when noise rate $\eta_{y,k} < 1 - \eta_y$ with $\sum_{i \neq y} \eta_{y,i} = \eta_y$, then*

$$R_{\mathcal{S}}^{\eta}(f^*) - R_{\mathcal{S}}^{\eta}(f) \simeq (1 - \frac{\eta K}{K-1})(R_{\mathcal{S}}(f^*) - R_{\mathcal{S}}(f)) \leq 0$$

From Theorem 2, we can derive that our method is nearly noise-tolerant under symmetric noise and further prove the robustness of the proposed method to asymmetric noise. More details can be found in Appendix B. We show a significant improvement under noise regimes empirically in experiments.

4. Experiments

Training and evaluation setups. We conduct extensive experiments on the benchmark datasets, CIFAR-10/100. We set the perturbation budget to $\epsilon = 8/255$ under the ℓ_∞ norm-bounded constraint. We use ResNet-18 [15] as our default network architecture unless otherwise specified. For all experiments, the model is trained using SGD with a momentum of 0.9, weight decay of 5×10^{-4}, and an initial learning rate of 0.1 for a total of 200 epochs. The learning rate is decayed by a factor of 0.1 at the 100-th and 150-th epochs, following [22]. The evaluation of the proposed approach encompasses PGD-20 [22] and AutoAttack [7], which is recognized as the most reliable robustness evaluation up to date. AutoAttack is an ensemble of diverse attacks, including APGD-CE, APGD-DLR, FAB [8] and Square attack [2]. We quantify the robust generalization by computing the difference between the best and final checkpoints over the course of training.

Improved robust performance across datasets. Experimental results of PGD-AT [18], TRADES [39], and the combinations of them with our proposal (denoted as PGD-AT+SGLR and TRADES+SGLR) on CIFAR-10/100 datasets are shown in Tab. 3, from which we make the following observations on our advantages: 1) **Closer generalization gap**. It is remarkable that the differences between the best and the final test accuracies of the combinations are reduced to around 0.5%, while the corresponding baselines (PGD-AT and TRADES) induce much larger gaps, which

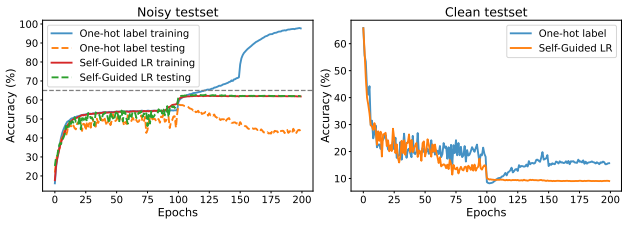


Figure 4. Test accuracy (%) on CIFAR-10 dataset (with 40% label noise). We split the training set into 1) *untouched portion*, where the labels of elements are left untouched; 2) *corrupted portion*, where the labels of elements are assigned uniformly at random.

Table 2. 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

Table 3. 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**

are up to 9.3% and 3.0% on CIFAR-10 dataset when the model is threaten by PGD-20. It indicates that our method effectively alleviates robust overfitting. 2) **Higher robust accuracy**. As the combinations induce smaller gaps, they can approach higher robust accuracy against adversaries compared to baseline methods, as can be observed in Tab. 3. Besides, it is notable that our method is especially effective against AutoAttack [7], regarding the fundamental difficulties of simultaneously achieving the robustness against multiple adversaries. 3) **Consistent improvement**. Combining our label smoothing method could consistently and significantly improve the performance on all considered adversaries and also on the clean test set, and across different datasets, namely, the CIFAR-10 and CIFAR-100 datasets.

Improved generalization under noise regimes. Moreover, we verify the self-guided label smoothing by comparing the behavior of the proposed SGLR w ith that of the widely used one-hot labels under different noise settings, with 40% noisy labels and fully 60% true labels, respectively. Specifically, we focus on evaluating the effectiveness of our approach in mitigating the impact of *symmetric* noisy labels, where the labels are resampled from a uniform distribution over all labels with a probability η.

We begin by making the observations in the left of Fig. 4: 1) The generalization error (*i.e.*, the difference between the training and test accuracy) of the proposed SGLR (almost negligible) is obviously smaller than that of the commonly used hard labels ($> 50\%$). 2) The peak of the training accuracy using hard labels (blue curve) can approximate 100% even on the training set with 40% noisy labels (horizontal gray dashed denotes the portion of correct labels), which suggests that using hard labels could fit correct labels in the early stage and eventually memorizes noisy labels. On the contrary, our training curve is bounded by the untouched portion, implying that our method is able to calibrate the training process and maintain a proper fitting of training data. Then we turn to the other setting as shown in the right of Fig. 4, our test accuracy grows steadily while the baseline method fails. This observation again suggests that the double-descent phenomenon might be due to the overfitting of noise and it is hopeful to be avoided by our method.

To give a more intuitive explanation, we further visualize the penultimate layer representations of ResNet 18 trained with (a) a hard label; (b) a soft label, and (c) the proposed soften label. As Fig. 5 shown, under all the settings of noise rate, the proposed SGLR with self-guided distribution

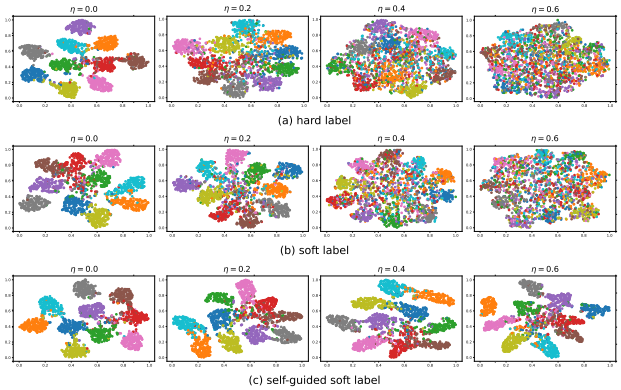


Figure 5. Visualization of representations learned by standard training with hard/soft labels and the proposed SGLR with self-guided distribution on CIFAR-10 dataset under various levels of symmetric noisy labels ($\eta \in [0.0, 0.2, 0.4, 0.6]$).

can consistently provide better representations, displayed with more separable clusters and tighter intra-class distance, which are very close to those of the clean setting (the leftmost column). Notably, even with a significantly increased noise rate from 0 (no noise) to 0.6 (severe noise), as illustrated from left to right in Fig. 5, our method learns representations with almost negligible variance. This also echoes our advantages of robust learning under various settings of data corruption.

In a nutshell, we have grounds to empirically believe that our method exhibits benign tolerance to noisy labels and improves the generalization ability of the model by avoiding fitting noisy labels.

Comparison against other methods. Considering that our method is analogous to employing the technique of label smoothing into AT, in Tab. 2, we report robustness evaluations of AT with label smoothing (PGD-AT + LS) and temporal ensembling (PGD-AT + TE) on CIFAR-10 test set. Since excessive LS could degrade the robustness reported in [20], we implement LS with smoothing level $r = 0.2$. We empirically found that SGLR not only can effectively alleviate robust overfitting while LS fails but also boosts the robustness even under strong AutoAttack. Further, we also report other methods (*e.g.*, AWP [33], KD-AT [10]) in Tab. 2, which can mitigate robust overfitting through the lens of weight smoothness and the knowledge transfer of teacher model. Though robust overfitting is indeed impeded by applying these methods, our method could better narrow the generalization gap between best and final models and achieve remarkable robustness than others under AutoAttack.

On the impact of smoothing effect r. The effectiveness of SGLR relies on the smoothing level of soft labels. Besides, as we discussed in Sec. 3.3, SGLR could be viewed as a kind of knowledge distillation without extra teacher models involved. Large temperature T during distillation improves the smoothness of output distribution but could impair the test performance. As the temperature is vital, we could specify $f(x; w)_i$ in Eq. (7) as $\exp(z_i/T)/\sum_j \exp(z_j/T)$. We vary the smoothing level $r \in \{0.0, 0.2, 0.4, 0.6, 0.8\}$ in Tab. 4 and note that an increase in r initially boosts both standard accuracy and robust accuracy, followed by a decline, indicating that excessive smoothing introduces noise and hampers predictive capability. Additionally, increasing the temperature while fixing r only gains slightly improvement in standard accuracy but oberves degradation in robust accuracy.

Table 4. Ablation study on smoothing level r and temperature T.

T/r		0.0	0.2	0.4	0.6	0.8
1	SA	82.4	83.5	82.8	78.5	69.1
	RA	43.7	54.9	54.2	52.7	48.3
1.5	SA	83.2	82.9	83.1	80.6	74.5
	RA	45.7	55.9	53.2	50.4	49.8
2	SA	82.9	84.4	83.9	78.5	75.6
	RA	45.8	53.9	52.0	50.9	47.8

5. Discussion and conclusion

In this study, we show that label noise induced by distribution mismatch and improper label assignments would degrade the test accuracies as well as make the robust overfitting aggravated. From the view of this observation, a label assignment approch for AT, Self-Guided Label Refinement (SGLR), is proposed to weaken the memorization in AT on noisy labels and thus to mitigate the robust overfitting. Extensive experimental results demonstrate the effectiveness of the proposed SGLR. Though we circumvent over confident prediction, the model is still not well-calibrated. As we measure the calibration of model during the training over the entire dataset instead of in a sample-wise way, it may give a false sense of reweighting confidence. In the future, we will attempt to address this limitation.

Acknowledgement

This work is partially supported by the National Science and Technology Major Project of China (Grant No. 2022ZD0119103), the National Natural Science Foundation of China (Grant No. 61906168), the Zhejiang Provincial Natural Science Foundation of China (Grant No.LQ21F020006, LY23F020023), and is also supported by the advanced computing resources provided by the Supercomputing Center of Hangzhou City University.

References

[1] Alessandro Achille and Stefano Soatto. Emergence of invariance and disentanglement in deep representations. *J. Mach. Learn. Res.*, 19:50:1–50:34, 2018. 3

[2] Maksym Andriushchenko, Francesco Croce, Nicolas Flammarion, and Matthias Hein. Square attack: A query-efficient black-box adversarial attack via random search. In *ECCV*, pages 484–501. Springer, 2020. 6

[3] Anish Athalye, Nicholas Carlini, and David A. Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *ICML*, pages 274–283. PMLR, 2018. 1

[4] Nicholas Carlini, Anish Athalye, Nicolas Papernot, Wieland Brendel, Jonas Rauber, Dimitris Tsipras, Ian J. Goodfellow, Aleksander Madry, and Alexey Kurakin. On evaluating adversarial robustness. *CoRR*, abs/1902.06705, 2019. 4

[5] Tianlong Chen, Zhenyu Zhang, Sijia Liu, Shiyu Chang, and Zhangyang Wang. Robust overfitting may be mitigated by properly learned smoothening. In *ICLR*. OpenReview.net, 2021. 1, 5, 6

[6] Hao Cheng, Zhaowei Zhu, Xing Sun, and Yang Liu. Mitigating memorization of noisy labels via regularization between representations. In *ICLR*. OpenReview.net, 2023. 2

[7] Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *ICML*, pages 2206–2216. PMLR, 2020. 6, 7

[8] Francesco Croce and Matthias Hein. Minimally distorted adversarial examples with a fast adaptive boundary attack. In *ICML*, pages 2196–2205. PMLR, 2020. 6

[9] Terrance Devries and Graham W. Taylor. Improved regularization of convolutional neural networks with cutout. *CoRR*, abs/1708.04552, 2017. 1

[10] Chengyu Dong, Liyuan Liu, and Jingbo Shang. Label noise in adversarial training: A novel perspective to study robust overfitting. In *NeurIPS*, 2022. 1, 2, 8

[11] Yu Feng and Yuhai Tu. Phases of learning dynamics in artificial neural networks in the absence or presence of mislabeled data. *Mach. Learn. Sci. Technol.*, 2(4):43001, 2021. 3

[12] Aritra Ghosh, Himanshu Kumar, and P. S. Sastry. Robust loss functions under label noise for deep neural networks. In *AAAI, February 4-9, San Francisco, California, USA*, pages 1919–1925. AAAI Press, 2017. 6

[13] Micah Goldblum, Liam Fowl, Soheil Feizi, and Tom Goldstein. Adversarially robust distillation. In *AAAI*, pages 3996–4003. AAAI Press, 2020. 5, 6

[14] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *ICLR*, 2015. 1

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778. IEEE Computer Society, 2016. 6

[16] Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. Distilling the knowledge in a neural network. *CoRR*, abs/1503.02531, 2015. 5

[17] Michal Lukasik, Srinadh Bhojanapalli, Aditya Krishna Menon, and Sanjiv Kumar. Does label smoothing mitigate label noise? In *ICML*, pages 6448–6458. PMLR, 2020. 4

[18] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *ICLR*. OpenReview.net, 2018. 1, 2, 6

[19] Mahdi Pakdaman Naeini, Gregory F. Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using bayesian binning. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, January 25-30, 2015, Austin, Texas, USA*, pages 2901–2907. AAAI Press, 2015. 3

[20] Tianyu Pang, Xiao Yang, Yinpeng Dong, Hang Su, and Jun Zhu. Bag of tricks for adversarial training. In *ICLR*. OpenReview.net, 2021. 4, 8

[21] Gabriel Pereyra, George Tucker, Jan Chorowski, Lukasz Kaiser, and Geoffrey E. Hinton. Regularizing neural networks by penalizing confident output distributions. In *ICLR*. OpenReview.net, 2017. 4

[22] Leslie Rice, Eric Wong, and J. Zico Kolter. Overfitting in adversarially robust deep learning. In *ICML*, pages 8093–8104. PMLR, 2020. 1, 2, 6, 4

[23] Amartya Sanyal, Puneet K. Dokania, Varun Kanade, and Philip H. S. Torr. How benign is benign overfitting ? In *ICLR*. OpenReview.net, 2021. 5, 3

[24] Gaurang Sriramanan, Sravanti Addepalli, Arya Baburaj, and Venkatesh Babu R. Towards efficient and effective adversarial training. In *NeurIPS*, pages 11821–11833, 2021. 1

[25] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *ICLR*, 2014. 1

[26] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *CVPR*, pages 2818–2826. IEEE Computer Society, 2016. 2

[27] Qizhou Wang, Feng Liu, Bo Han, Tongliang Liu, Chen Gong, Gang Niu, Mingyuan Zhou, and Masashi Sugiyama. Probabilistic margins for instance reweighting in adversarial training. In *NeurIPS*, pages 23258–23269, 2021. 1

[28] Yisen Wang, Xingjun Ma, Zaiyi Chen, Yuan Luo, Jinfeng Yi, and James Bailey. Symmetric cross entropy for robust learning with noisy labels. In *ICCV*, pages 322–330. IEEE, 2019. 5

[29] Yisen Wang, Difan Zou, Jinfeng Yi, James Bailey, Xingjun Ma, and Quanquan Gu. Improving adversarial robustness requires revisiting misclassified examples. In *ICLR*. OpenReview.net, 2020. 1, 5

[30] Yifei Wang, Liangchen Li, Jiansheng Yang, Zhouchen Lin, and Yisen Wang. Balance, imbalance, and rebalance: Understanding robust overfitting from a minimax game perspective. In *NeurIPS*, 2023. 2

[31] Ziqiao Wang and Yongyi Mao. On the generalization of models trained with SGD: information-theoretic bounds and implications. In *ICLR*. OpenReview.net, 2022. 3

[32] Zifeng Wang, Shao-Lun Huang, Ercan Engin Kuruoglu, Jimeng Sun, Xi Chen, and Yefeng Zheng. Pac-bayes information bottleneck. In *ICLR*. OpenReview.net, 2022. 4

[33] Dongxian Wu, Shu-Tao Xia, and Yisen Wang. Adversarial weight perturbation helps robust generalization. In *NeurIPS*, 2020. 1, 8

[34] Xiaobo Xia, Tongliang Liu, Bo Han, Chen Gong, Nannan Wang, Zongyuan Ge, and Yi Chang. Robust early-learning: Hindering the memorization of noisy labels. In *ICLR*. OpenReview.net, 2021. 2

[35] Zitong Yang, Yaodong Yu, Chong You, Jacob Steinhardt, and Yi Ma. Rethinking bias-variance trade-off for generalization of neural networks. In *ICML*, pages 10767–10777. PMLR, 2020. 1

[36] Chaojian Yu, Bo Han, Mingming Gong, Li Shen, Shiming Ge, Du Bo, and Tongliang Liu. Robust weight perturbation for adversarial training. In *IJCAI*, pages 3688–3694. ijcai.org, 2022. 1

[37] Chaojian Yu, Bo Han, Li Shen, Jun Yu, Chen Gong, Mingming Gong, and Tongliang Liu. Understanding robust overfitting of adversarial training and beyond. In *ICML*, pages 25595–25610. PMLR, 2022. 1, 2

[38] Hongyi Zhang, Moustapha Cissé, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *ICLR*. OpenReview.net, 2018. 1

[39] Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric P. Xing, Laurent El Ghaoui, and Michael I. Jordan. Theoretically principled trade-off between robustness and accuracy. In *ICML*, pages 7472–7482. PMLR, 2019. 1, 6

[40] Jingfeng Zhang, Jianing Zhu, Gang Niu, Bo Han, Masashi Sugiyama, and Mohan S. Kankanhalli. Geometry-aware instance-reweighted adversarial training. In *ICLR*. OpenReview.net, 2021. 1

[41] Bojia Zi, Shihao Zhao, Xingjun Ma, and Yu-Gang Jiang. Revisiting adversarial robustness distillation: Robust soft labels make student better. In *ICCV*, pages 16423–16432. IEEE, 2021. 5, 6