RELIABLE ADVERSARIAL DISTILLATION WITH UNRELIABLE TEACHERS

Anonymous authors
Paper under double-blind review

ABSTRACT

In ordinary distillation, student networks are trained with *soft labels* (SLs) given by *pretrained* teacher networks, and students are expected to improve upon teachers since SLs are *stronger* supervision than the original *hard labels*. However, when considering *adversarial robustness*, teachers may become unreliable and adversarial distillation may not work: teachers are pretrained on their own adversarial data, and it is too demanding to require that teachers are also good at every adversarial data queried by students. Therefore, in this paper, we propose reliable *introspective adversarial distillation* (IAD) where students *partially* instead of *fully* trust their teachers. Specifically, IAD distinguishes between three cases given a query of a *natural data* (ND) and the corresponding *adversarial data* (AD): (a) if a teacher is good at AD, its SL is fully trusted; (b) if a teacher is good at ND but not AD, its SL is partially trusted and the student also takes its own SL into account; (c) otherwise, the student only relies on its own SL. Experiments demonstrate the effectiveness of IAD for improving upon teachers in terms of adversarial robustness.

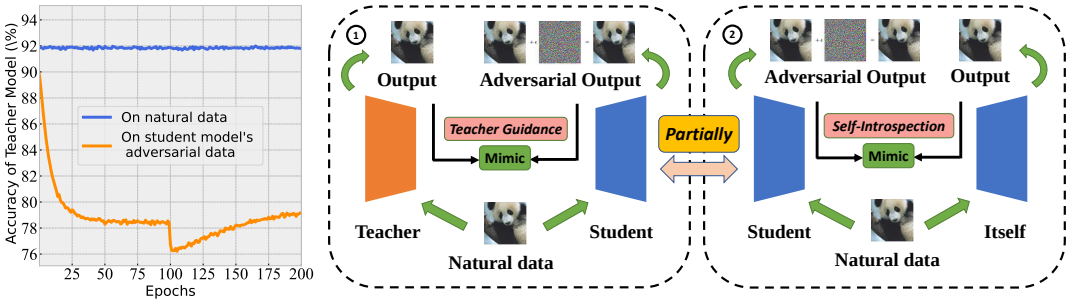
1 INTRODUCTION

Deep Neural Networks (DNNs) have shown excellent performance on a range of tasks in computer vision (He et al., 2016) and natural language processing (Devlin et al., 2019). Nevertheless, Szegedy et al. (2014); Goodfellow et al. (2015) demonstrated that DNNs could be easily fooled by adding a small number of perturbations on natural examples, which increases the concerns on the robustness of DNNs in the trustworthy-sensitive areas, *e.g.,* finance (Kumar et al., 2020) and autonomous driving (Litman, 2017). To overcome this problem, adversarial training (Goodfellow et al., 2015; Madry et al., 2018) is proposed and has shown effectiveness to acquire the adversarially robust DNNs.

Most existing adversarial training approaches focus on learning from data directly. For example, the popular adversarial training (AT) (Madry et al., 2018) leverages multi-step projected gradient descent (PGD) to generate the adversarial examples and feed them into the standard training. Zhang et al. (2019) developed TRADES on the basis of AT to balance the standard accuracy and robust performance. Recently, there are several methods under this paradigm are developed to improve the model robustness (Wang et al., 2019; Alayrac et al., 2019; Zhang et al., 2020; Jiang et al., 2020; Ding et al., 2020; Wang et al., 2020b; Du et al., 2021; Tian et al., 2021; Zhang et al., 2021). However, directly learning from the adversarial examples is a challenging task on the complex datasets since the loss with hard labels is difficult to be optimized, which limits us to achieve higher robust accuracy.

To mitigate this issue, one emerging direction is distilling robustness from the adversarially pre-trained model intermediately, which has shown promise in the recent study. For example, Ilyas et al. (2019) used an adversarially pre-trained model to build a "robustified" dataset to learn a robust DNN. Chen et al. (2020); Salman et al. (2020) explored to boost the model robustness through fine-tuning or transfer learning from adversarially pre-trained models. Goldblum et al. (2020)and Chen et al. (2021) investigated distilling the robustness from adversarially pre-trained models, termed as *adversarial distillation* for simplicity, where they encouraged student models to mimic the outputs (i.e., soft labels) of the adversarially pre-trained teachers.

However, one critical difference is: in the conventional distillation, the teacher model and the student model share the natural training data; while in the adversarial distillation, the adversarial training data of the student model and that of the teacher model are egocentric (respectively generated by themselves) and becoming more adversarial challenging during training. Given this distinction, are the



(a) Unreliability issue

(b) Overview of Introspective Adversarial Distillation (IAD)

Figure 1: **(a) Unreliability issue**: Comparison of the teacher model's accuracy on the natural data and the student model's adversarial data from *CIFAR-10* datasets. Different from the ordinary distillation in which the natural data is unchanged and the teacher model has the consistent standard performance, the teacher model's robust accuracy on the student model's adversarial data is decreasing during distillation, which means the guidance of the teacher model in adversarial distillation is progressively unreliable. **(b) Overview of Introspective Adversarial Distillation (IAD)**: The student *partially trusts the teacher guidance* and *partially trusts self-introspection* during adversarial distillation. Specifically, the student model generates the adversarial data by itself and mimics the outputs of the teacher model and itself partially in IAD.

soft labels acquired from the teacher model in adversarial distillation always reliable and informative guidance? To answer this question, we take a closer look at the process of adversarial distillation. As shown in Figure 1(a), we discover that along with the training, the teacher model progressively fails to give a correct prediction for the adversarial data queried by the student model. The reason could be that with the students being more adversarially robust and thus the adversarial data being harder, it is too demanding to require the teachers become always good at every adversarial data queried by the student model, as the teacher model has never seen these data in its pre-training. In contrast, for the conventional distillation, student models are expected to distill the *"static"* knowledge from the teacher model, since the soft labels for the natural data from the teacher model are always fixed.

The observation in Figure 1(a) raises the challenge: *how to conduct reliable adversarial distillation with unreliable teachers*? To solve this problem, we can categorize the training data according to the prediction on natural and student generated adversarial data into three cases. First, if the teacher model can correctly classify both natural and adversarial data, it is reliable; Second, if the teacher model can correctly classify the natural but not adversarial data, it should be partially trusted, and the student model is suggested to trust itself to enhance model robustness as the adversarial regularization (Zhang et al., 2019); Third, if the teacher model cannot correctly classify both natural and adversarial data, the student model is recommended to trust itself totally. According to this intuition, we propose an Introspective Adversarial Distillation (IAD) to effectively utilize the knowledge from an adversarially pre-trained teacher model. The framework of our proposed IAD can be seen in Figure 1(b). Briefly, the student model is encouraged to *partially* instead of *fully* trust the teacher model, and gradually trust itself more as being more adversarially robust. We conduct extensive experiments on the benchmark *CIFAR-10/CIFAR-100* and the more challenging *Tiny-ImageNet* datasets to evaluate the efficiency of our IAD. The main contributions of our work can be summarized as follows.

1. We take a closer look at adversarial distillation under the teacher-student paradigm. Considering adversarial robustness, we discover that the guidance from the teacher model is progressively unreliable along with the adversarial training.
2. We construct the reliable guidance for adversarial distillation by flexibly utilizing the robust knowledge from the teacher model: (a) if a teacher is good at adversarial data, its soft labels can be fully trusted; (b) if a teacher is good at natural data but not adversarial data, its soft labels should be partially trusted and the student also takes its own soft labels into account; (c) otherwise, the student only relies on its own soft labels.
3. We propose an Introspective Adversarial Distillation (IAD) to automatically realize the intuition of the previous reliable guidance during the adversarial distillation. The experimental results confirmed that our approach can improve adversarial robustness across a variety of training settings and evaluations, especially on the challenging (consider adversarial

robustness) datasets (e.g., *CIFAR-100* (Krizhevsky, 2009) and *Tiny-ImageNet* (Le & Yang, 2015)) or using large models (e.g., *WideResNet* (Zagoruyko & Komodakis, 2016)).

2 Related Work

2.1 Adversarial Training.

Adversarial examples (Goodfellow et al., 2015) motivate many defensive approaches developed in the last few years. Among them, adversarial training has been demonstrated as the most effective method to improve the robustness of DNNs (Cai et al., 2018; Wang et al., 2020a;b; Jiang et al., 2020; Bai et al., 2021; Chen et al., 2021; Wu et al., 2020). The formulation of the popular AT (Madry et al., 2018) and its variants can be summarized as the minimization of the following loss:

$$\min_{f_\theta \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \ell(f_\theta(\tilde{x}_i), y_i), \tag{1}$$

where n is the number of training examples, $\tilde{x}_i$ is the adversarial example within the ϵ-ball (bounded by an L_p-norm) centered at natural example x_i, y_i is the associated label, f_θ is the DNN with parameter θ and $\ell(\cdot)$ is the standard classification loss, e.g., the cross-entropy loss. Adversarial training leverages adversarial examples to smooth the small neighborhood, making the model prediction locally invariant. To generate the adversarial examples, AT employs a PGD method (Madry et al., 2018). Concretely, given a sample $x^{(0)} \in \mathcal{X}$ and the step size $\beta > 0$, PGD recursively searches

$$\tilde{x}^{(t+1)} = \Pi_{\mathcal{B}[\tilde{x}^{(0)}]}\big(\tilde{x}^{(t)} + \beta\, \mathrm{sign}(\nabla_{\tilde{x}^{(t)}} \ell(f_\theta(\tilde{x}^{(t)}), y))\big), \tag{2}$$

until a certain stopping criterion is satisfied. In Eq. equation 2, $t \in \mathbb{N}$, ℓ is the loss function, $\tilde{x}^{(t)}$ is adversarial data at step t, y is the corresponding label for natural data, and $\Pi_{\mathcal{B}_\epsilon[\tilde{x}_0]}(\cdot)$ is the projection function that projects the adversarial data back into the ϵ-ball centered at $\tilde{x}^{(0)}$.

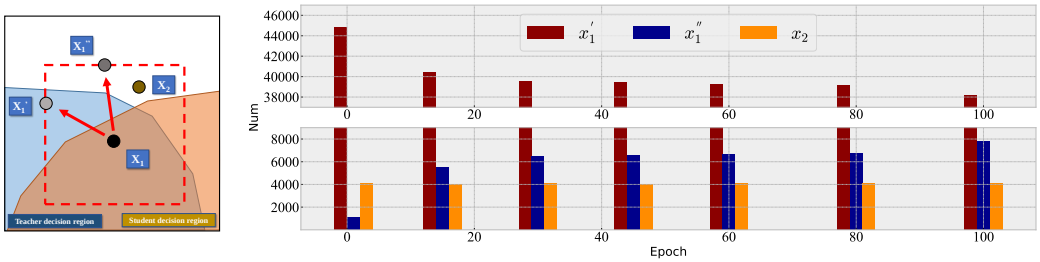
2.2 Knowledge Distillation

The idea of distillation from other models can be dated back to (Craven & Shavlik, 1996), and re-introduced by (Hinton et al., 2015) as knowledge distillation (KD). It has been widely studied in recent years and works well in numerous applications like model compression and transfer learning. For adversarial defense, a few studies have explored obtaining adversarial robust models by distillation. Papernot et al. (2016) proposed defensive distillation which utilizes the soft labels produced by a standard pre-trained teacher model, while this method is proved to be not resistant to the C&W attacks (Carlini & Wagner, 2017); Goldblum et al. (2020) combined AT with KD to transfer robustness to student models, and they found that the distilled models can outperform adversarially pre-trained teacher models of identical architecture in terms of adversarial robustness; Chen et al. (2021) utilized distillation as a regularization for adversarial training, which employed both robust and standard pre-trained teacher models to address the robust overfitting (Rice et al., 2020).

Nonetheless, all these related methods fully trust teacher models and do not consider that whether the guidance of the teacher model in distillation is reliable or not. In this paper, different from the previous studies, we find that the teacher model in adversarial distillation is not always trustworthy. Based on that, we propose reliable IAD to encourage student models to partially instead of fully trust teacher models, which effectively utilizes the knowledge from the adversarially pre-trained models.

3 A Closer Look at Adversarial Distillation

In Section 3.1, we discuss the unreliable issue of adversarial distillation, i.e., the guidance of the teacher model is progressively unreliable along with adversarial training. In Section 3.2, we partition the training examples into three parts and analyze them part by part. Specifically, we expect that the student model should *partially* instead of *fully* trust the teacher model and gradually trust itself more along with adversarial training.



(a) Toy illustration

(b) Number changes on the three kinds of data during distillation

Figure 2: **(a) Toy illustration**: An illustration of three situations about the prediction of the teacher model. The blue/orange areas represent the decision region (in which the model can give the correct predictions) of the teacher/student model, and the red dashed box represents the unit-norm ball of AT. 1) x_1'. The generated adversarial example from the natural example x_1 is located in the blue area, where the teacher model can correctly predict; 2) x_1''. The generated adversarial example from x_1 is out of both orange and blue areas, where the teacher model has the wrong prediction; 3) x_2. The natural example that both two models cannot correctly predict. **(b) Number changes on the three kinds of data**: We trace the number of three types of data during adversarial distillation on *CIFAR-10* dataset. Noted that, different from the consistent number of examples like x_2, the number of examples like x_1' is *decreasing* and that of examples like x_1'' is *increasing* during adversarial distillation.

3.1 Fully Trust: Progressively Unreliable Guidance

As aforementioned in the Introduction, previous methods (Goldblum et al., 2020; Chen et al., 2021) fully trust the teacher model when distilling robustness from adversarially pre-trained models. Taking Adversarial Robust Distillation (ARD) (Goldblum et al., 2020) as an example, we illustrate its procedure in the left part of Figure 1(b): the student model generates its adversarial data and then optimizes the prediction of them to mimic the output of the teacher model. However, although the teacher model is well optimized on the adversarial data queried by itself, we argue that it might not always be good at the more and more challenging adversarial data queried by the student model.

As shown in Figure 1(a), different from the ordinary distillation in which the teacher model has the consistent standard performance on the natural data, its robust accuracy on the student model's adversarial data is decreasing during distillation. The guidance of the teacher model gradually fails to give the correct output on the adversarial data queried by the student model.

3.2 Partially Trust: Construction of Reliable Guidance

The unreliable issue of the teacher model in adversarial distillation raises the challenge of how to conduct *reliable adversarial distillation with unreliable teachers*? Intuitively, this requires us to re-consider the guidance of adversarially pre-trained models along with the adversarial training. For simplicity, we use $\mathcal{T}(x)$ ($\mathcal{T}(\tilde{x})$) to represent the predicted label of the teacher model on the natural (adversarial) examples, and use y to represent the targeted label. We partition the adversarial samples into three parts as shown in the toy illustration (Figure 2(a)), and analyze them part by part.

1) $\mathcal{T}(x) = y \cap \mathcal{T}(\tilde{x}) = y$: It can be seen in Figure 2(a) that this part of data whose adversarial variants like x_1' is the most trustworthy among the three parts, since the teacher model performs well on both natural and adversarial data. In this case, we could choose to trust the guidance of the teacher model on this part of the data. However, as shown in Figure 2(b), we find that the sample number of this part is decreasing along with the adversarial training. That is, what we can rely on from the teacher model in adversarial distillation is progressively reduced.

2) $\mathcal{T}(x) = y \cap \mathcal{T}(\tilde{x}) \neq y$: In Figure 2(b), we also check the number change of the part of data whose adversarial variants like x_1''. Corresponding to the previous category, the number of this kind of data is increasing during distillation. Since the teacher model's outputs on the small neighborhood of the queried natural data are not always correct, its knowledge may not be robust and the guidance for the student model is not reliable. Think back to the reason for the decrease in the robust accuracy of the teacher model, the student model itself may also be trustworthy since it becomes gradually adversarial robust during distillation.

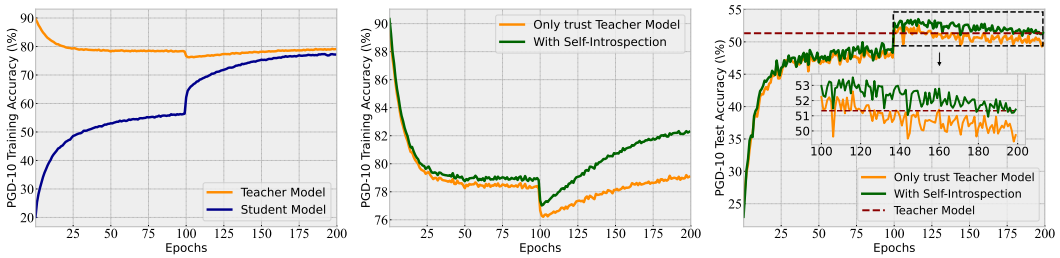


Figure 3: Reliability certification about self-introspection of the student model: *Left*, PGD-10 training accuracy of teacher/student model; *Middle*, PGD-10 training accuracy of teacher model and that combined with the self-introspection of the student model; *Right*, PGD-10 test accuracy of the student model which only trusts the soft labels and the student model which also considers self-introspection.

3) $\mathcal{T}(x) \neq y \cap \mathcal{T}(\tilde{x}) \neq y$: As for the data which are like x_2 in Figure 2(a), the guidance of the teacher model is totally unreliable since the predicted labels on the natural data are wrong. The student model may also trust itself to encourage the outputs to mimic that of their natural data rather than the wrong outputs from the teacher model. First, it removes the potential threat that the teacher's guidance may be a kind of noisy labels for training. Second, as an adversarial regularization (Zhang et al., 2019), it can improve the model robustness through enhancing the stability of the model's outputs on the natural and the corresponding adversarial data.

To sum up, we suggest employing reliable guidance from the teacher model and encouraging the student model to trust itself more as the teacher model's guidance being progressively unreliable and the student model gradually becoming more adversarially robust.

4 Introspective Adversarial Distillation

Based on previous analysis about the adversarial distillation, we propose the Introspective Adversarial Distillation (IAD) to better utilize the guidance from the adversarially pre-trained model. Concretely, we have the following KD-style loss, but composite with teacher guidance and student introspection.

$$\begin{aligned}\ell_{IAD} =& \lambda \ell_{CE}(S(x), y) + (1-\lambda) \cdot \tau^2 \cdot \\ & (\alpha \underbrace{\ell_{KL}(S(\tilde{x}|\tau)||T(x|\tau))}_{\textbf{Teacher Guidance}} + (1-\alpha) \underbrace{\ell_{KL}(S(\tilde{x}|\tau)||S(x|\tau))}_{\textbf{Student Introspection}}),\end{aligned} \tag{3}$$

where $S(\cdot|\tau)$ is the tempered variant of the student output $S(\cdot)$ with the temperature $\tau \in (0, +\infty)$, $T(\cdot|\tau)$ is the tempered variant of the teacher output $T(\cdot)$, $\tilde{x}$ is the adversarial data generated from the natural data x, y is the hard label, ℓ_{CE} is the Cross Entropy loss and ℓ_{KL} is the KL-divergence loss. As for the annealing parameter $\alpha \in [0, 1]$ that is used to balance the effect of the teacher model in adversarial distillation, based on the analysis about the reliability of adversarial supervision in Section 3, we define it as,

$$\alpha = (P_T(\tilde{x}|y))^{\beta}, \tag{4}$$

where $P_T(\cdot|y)$ is the prediction probability of the teacher model about the targeted label y and β is a hyperparameter to sharpen the prediction. The intuition behind IAD is to calibrate the guidance from the teacher model automatically based on the prediction of adversarial data. Our α naturally corresponds to the construction in Section 3.2, since the prediction probability of the teacher model for the adversarial data can well represent the categorical information. As for β, we have plot the specific values of α under its adjustment in the left of Figure 4.

Intuitively, the student model can trust the teacher model when α approaches 1, which means that the teacher model is good at both natural and adversarial data. However, when α approaches 0, it corresponds that the teacher model is good at natural but not adversarial data, or even not good at both, and thus the student model should take its self-introspection into account. In Figure 3, we check the reliability of the student model itself. According to the left panel of Figure 3, we can see that the student model is progressively robust to the adversarial data. And if we incorporate the student introspection into the adversarial distillation, the results in the middle of Figure 3 confirms its potential benefits to improve the accuracy of the guidance. Moreover, as shown in the right panel of

Algorithm 1 Introspective Adversarial Distillation (IAD)

Input: student model S, teacher model T, training dataset $D = \{(x_i, y_i)\}_{i=1}^{n}$, learning rate η, number of epochs N, batch size m, number of batches M, temperature parameter τ, teacher model's predicted probability α, adjustable parameter λ and β.
Output: adversarially robust model S_r
for epoch $= 1, \ldots, N$ **do**
  for mini-batch $= 1, \ldots, M$ **do**
    Sample a mini-batch $\{(x_i, y_i)\}_{i=1}^{m}$ from D
    for $i = 1, \ldots, m$ (in parallel) **do**
      Obtain adversarial data $\tilde{x}_i$ of x_i by PGD based on Eq. equation 2.
      Compute α for each adversarial data based on Eq. equation 4.
    end for
    $$\theta \leftarrow \theta - \eta \nabla_\theta \left\{ \begin{array}{l} \lambda \ell_{CE}(S(x), y) + (1-\lambda) \cdot \tau^2 \cdot \\ (\alpha \underbrace{\ell_{KL}(S(\tilde{x}_i|\tau)||T(x_i|\tau))}_{\textbf{Teacher Guidance}} + (1-\alpha) \underbrace{\ell_{KL}(S(\tilde{x}_i|\tau)||S(x|\tau))}_{\textbf{Student Introspection}}) \end{array} \right\}$$
  end for
end for

Figure 3, adding self-introspection results in better improvement in model robustness compared to only using the guidance of the teacher model. Therefore, ℓ_{IAD} automatically encourages the outputs of the student model to mimic more reliable guidance in adversarial distillation.

Algorithm 2 summarizes the implementation of Introspective Adversarial Distillation (IAD). Specifically, IAD first leverages PGD to generate the adversarial data for the student model. Secondly, IAD computes the outputs of the teacher model and the student model on the natural data. Then, IAD mimics the outputs of the student model with that of itself and the teacher model partially based on the probability of the teacher model on the adversarial data.

Warming-up period. During training, we add a warming-up period to activate the student model, where α (in Eq. equation 3) is hardcoded to 1. This is because the student itself is not trustworthy in the early stage (refer to the left panel of Figure 3). Through that, we expect the student model to first evolve into a relatively reliable learner and then conducts the procedure of IAD.

4.1 Comparison with Related Methods

In this section, we discuss the difference between IAD and other related approaches in the perspective of the loss functions. Table 1 summarizes all of them.

Table 1: Loss comparison. Note that, $\lambda,\ \lambda_i \geq 0$, $\sum_i \lambda_i = 1$ and $\alpha = (P_T(\tilde{x}|y))^\beta$.

Method	Loss Function
AT	$\ell_{CE}(S(\tilde{x}), y)$
TRADES	$\ell_{CE}(S(x), y) + \lambda \ell_{KL}(S(\tilde{x})\|\|S(x))$
ARD	$\lambda \ell_{CE}(S(x), y) + (1-\lambda)\tau^2 \ell_{KL}(S(\tilde{x}\|\tau)\|\|T(x\|\tau))$
AKD^2	$\lambda_1 \ell_{CE}(S(\tilde{x}), y) + \lambda_2 \tau^2 \ell_{KL}(S(\tilde{x}\|\tau)\|\|T_{at}(\tilde{x}\|\tau)) + \lambda_3 \tau^2 \ell_{KL}(S(\tilde{x}\|\tau)\|\|T_{st}(\tilde{x}\|\tau)))$
IAD	$\lambda \ell_{CE}(S(x), y) + (1-\lambda)\tau^2(\alpha \ell_{KL}(S(\tilde{x}\|\tau)\|\|T(x\|\tau)) + (1-\alpha)\ell_{KL}(S(\tilde{x}\|\tau)\|\|S(x\|\tau)))$

As shown in Table 1, AT (Madry et al., 2018) utilizes the hard labels to supervise adversarial training; TRADES (Zhang et al., 2019) decomposes the loss function of AT into two terms, one for standard training and the other one for adversarial training with the soft supervision; Motivated by KD (Hinton et al., 2015), Goldblum et al. (2020) proposed ARD to conduct the adversarial distillation, which fully trusts the outputs of the teacher model to learn the student model. As indicated by the experiments in Goldblum et al. (2020), a larger λ resulted in less robust student models. Thus, they generally set $\lambda = 0$ in their experiments; Chen et al. (2021) utilized distillation as a regularization to avoid the robust overfitting issue, which employed both the adversarially pre-trained teacher model and the standard pre-trained model. Thus, there are two KL-divergence loss and for simplicity, we term their

Table 2: Test accuracy (%) on *CIFAR-10/CIFAR-100* datasets using *ResNet-18*.

Datasets	CIFAR-10					CIFAR-100				
Methods \ Measure	Natural	FGSM	PGD-20	CW_∞	AA	Natural	FGSM	PGD-20	CW_∞	AA
AT	83.06%	63.53%	50.21%	49.22%	46.70%	56.21%	33.94%	24.60%	23.35%	21.55%
ARD (AT)	83.13%	**64.11%**	51.36%	50.37%	48.05%	56.65%	35.82%	27.04%	25.30%	23.34%
AKD^2 (AT+ST)	**83.52%**	63.91%	51.36%	50.36%	48.08%	56.46%	35.76%	27.18%	25.33%	23.47%
IAD (AT)	83.33%	63.90%	51.77%	50.63%	**48.59%**	55.88%	35.68%	27.32%	25.60%	23.96%
AKD^2+**IAD** (AT+ST)	83.21%	63.54%	**51.85%**	**50.67%**	48.58%	**57.26%**	**35.94%**	**27.50%**	**25.68%**	**24.06%**
TRADES	81.26%	63.12%	52.98%	50.29%	49.47%	54.01%	35.23%	28.00%	24.26%	23.46%
ARD (TRADES)	81.50%	63.38%	53.38%	51.27%	49.92%	55.82%	**37.56%**	30.21%	26.76%	25.36%
AKD^2 (TRADES+ST)	83.49%	64.00%	51.89%	50.25%	48.52%	**57.46%**	37.32%	28.62%	25.76%	24.36%
IAD (TRADES)	80.62%	63.52%	**53.83%**	**51.41%**	**50.15%**	54.75%	37.19%	**30.71%**	**27.20%**	**26.10%**
AKD^2+**IAD** (TRADES+ST)	**83.76%**	**64.17%**	52.16%	50.59%	48.91%	57.08%	36.94%	29.08%	25.83%	24.45%

method as AKD^2; Regarding IAD, two KD-style losses respectively encourage the student model to partially instead of fully trust the guidance of the teacher model and gradually trust itself more. In the loss function, $\alpha = P_T(\tilde{x}|y)^\beta$ is gradually decreased as the adversarial examples being more challenging, which reduces dependency on the guidance of the teacher model during training.

5 EXPERIMENTS

We conduct extensive experiments to evaluate the effectiveness of IAD. In Section 5.1, we compare IAD with benchmark adversarial training methods (AT and TRADES) and some related methods which utilize adversarially pre-trained models via KD (ARD and AKD^2) on *CIFAR-10/CIFAR-100* (Krizhevsky, 2009) datasets. In Section 5.2, we compare the previous methods with IAD on a more challenging dataset *Tiny-ImageNet* (Le & Yang, 2015). In Section 5.3, the ablation studies are conducted to analyze the effects of the hyper-parameter β and different warming-up periods for IAD.

Performance evaluations. Several measures regarding both natural accuracy and adversarial robustness are applied to evaluate the model performance. We compute the natural accuracy on the natural test data and the robust accuracy on the adversarial test data following Wang et al. (2019). Specifically, the adversarial test data are generated by FGSM, PGD-20, and C&W attacks with the same perturbation bound $\epsilon = 8/255$ and the step size $\sigma = 2/255$. All the adversarial generation in these attacks has a random start, i.e, the uniformly random perturbation of $[-\epsilon, \epsilon]$ added to the natural data before attacking iterations. Moreover, we also estimate the model performance under the AutoAttack (termed as AA for simplicity).

5.1 EVALUATION ON *CIFAR-10/CIFAR-100* DATASETS

Experiment Setup. In this part, we follow the setup (learning rate, optimizer, weight decay, momentum) of Goldblum et al. (2020) to implement the adversarial distillation experiments on the *CIFAR-10/CIFAR-100* datasets. Specifically, we train ResNet-18 under different methods using SGD with 0.9 momentum for 200 epochs. The initial learning rate is 0.1 divided by 10 at Epoch 100 and Epoch 150 respectively, and the weight decay=0.0002. In the settings of adversarial training, we set the perturbation bound $\epsilon = 8/255$, the PGD step size $\sigma = 2/255$, and PGD step numbers $K = 10$. In the settings of distillation, we use $\tau = 1$ and use models pre-trained by AT and TRADES which have the best PGD-10 test accuracy as the teacher models for ARD, AKD^2 and our IAD. For ARD, we set its hyper-parameter $\lambda = 0$ as recommend in Goldblum et al. (2020) for gaining better robustness. For AKD^2, we set $\lambda_1 = 0.25$, $\lambda_2 = 0.5$ and $\lambda_3 = 0.25$ as recommanded in Chen et al. (2021). For IAD, we set $\lambda = 0, \beta = 0.1$ and warming-up period as 40/80 epoch to train on *CIFAR-10*/*CIFAR-100*. In addition, we also apply our IAD in AKD^2 to check its effectiveness on improving the robustness.

Results. We report the results in Table 2, where the results of AT and TARDES are listed in the first and fifth rows of Table 2, and the other methods use these models as the teacher models in distillation. On *CIFAR-10* dataset, we note that our IAD (or AKD^2+IAD) has obtained consistent improvements on adversarial robustness in terms of PGD-20, $CW\infty$ and AA accuracy compared with the student models distilled by ARD or AKD^2 and the adversarially pre-trained teacher models. While the natural and FGSM accuracy of our IAD is lower than others since we encourage the student model

Table 3: Test accuracy (%) on *CIFAR-10/CIFAR-100* datasets using *WideResNet-34-10*.

Datasets	CIFAR-10					CIFAR-100				
Methods \ Measure	Natural	FGSM	PGD-20	CW_∞	AA	Natural	FGSM	PGD-20	CW_∞	AA
AT	85.24%	66.07%	53.36%	52.85%	50.37%	59.75%	39.67%	31.14%	29.99%	27.46%
ARD (AT)	83.39%	64.43%	54.06%	53.27%	51.57%	58.26%	37.59%	30.03%	28.62%	26.36%
AKD^2 (AT+ST)	**86.25%**	**67.65%**	55.05%	54.07%	51.70%	**61.57%**	**41.14%**	31.90%	30.45%	27.80%
IAD (AT)	84.10%	66.24%	55.37%	54.20%	52.03%	58.19%	40.34%	**32.80%**	**31.03%**	**29.11%**
AKD^2+**IAD** (AT+ST)	85.09%	66.54%	**55.45%**	**54.63%**	**52.29%**	60.72%	40.67%	32.33%	30.40%	27.89%
TRADES	82.90%	65.55%	55.50%	53.04%	52.00%	59.14%	38.81%	31.21%	27.69%	26.53%
ARD (TRADES)	82.03%	64.67%	54.82%	53.23%	51.08%	58.63%	37.98%	29.08%	26.45%	25.92%
AKD^2 (TRADES+ST)	84.95%	67.14%	55.38%	53.38%	52.12%	**61.48%**	40.52%	32.34%	29.84%	27.84%
IAD (TRADES)	82.89%	65.58%	**55.86%**	53.63%	**52.29%**	58.32%	**40.82%**	**33.60%**	**30.13%**	**29.05%**
AKD^2+**IAD** (TRADES+ST)	**85.68%**	**67.39%**	55.45%	**53.77%**	51.85%	60.82%	40.43%	32.39%	30.01%	28.01%

Table 4: Test accuracy (%) on *Tiny-ImageNet* dataset using *PreActive-ResNet-18*.

Methods	Natural	FGSM	PGD-20	CW_∞	AA
AT	46.16%	28.20%	22.16%	19.52%	18.04%
ARD (AT)	47.34%	29.56%	22.60%	19.96%	17.60%
AKD^2 (AT+ST)	49.48%	30.10%	22.70%	20.10%	18.10%
IAD (AT)	46.66%	29.30%	22.68%	19.64%	17.82%
AKD^2+**IAD** (AT)	**49.52%**	**30.40%**	**23.42%**	**21.26%**	**18.64%**
TRADES	46.12%	27.74%	21.00%	16.86%	15.86%
ARD (TRADES)	**49.60%**	29.98%	22.76%	19.16%	17.26%
AKD^2 (TRADES+ST)	47.72%	28.12%	22.10%	18.28%	17.13%
IAD (TRADES)	47.12%	29.12%	23.28%	19.52%	**17.92%**
AKD^2+**IAD** (TRADES+ST)	48.10%	**30.54%**	**23.72%**	**19.78%**	17.82%

to partially trust itself which enhances the stability of model outputs but sacrifice a part of standard performance. On *CIFAR-100* dataset, the improvements of our IAD for adversarial robustness are more obvious compared with that on *CIFAR-10*, especially distilling from the teacher model trained by TRADES. However, since the teacher models have poor standard and robust performance on *CIFAR-100*, the models distilled by the ARD and AKD^2 which fully trust the guidance of teacher models also result in worse robust performance compared with our IAD.

Experiment Setup. In this part, we evaluate these methods by the model with larger capacity, i.e, *WideResNet-34-10*. For these adversarially pre-trained models, we follow the settings of (Zhang et al., 2021) to train AT and TRADES. To be specific, we train *WideResNet-34-10* using SGD with 0.9 momentum for 100 epochs. The initial learning rate is 0.1 divided by 10 at Epoch 30 and 60 respectively, and the weight decay=0.0005. For distillation baselines, we keep the most settings same as that in the previous part. Specially, we adjust the $\lambda = 0.95$ for ARD on *CIFAR-100* dataset as (Goldblum et al., 2020) recommend to deal with complex tasks. For IAD, we use $\lambda = 0.95, \beta = 0.1$ and 5/10 epochs warming-up period for *CIFAR-10*/*CIFAR-100*.

Results. We report the results in Tables 3. According the results on *CIFAR-10* dataset, our method can achieve better model robustness than ARD, AKD^2 and the original teacher models in terms of PGD-20, $CW\infty$ and AA accuracy. Moreover, our IAD does not sacrifice much standard performance (See the results of IAD (TRADES) and TRADES in Table 3). Since AKD^2 externally utilizes a standard pre-trained teacher model, it can achieve better natural accuracy in adversarial distillation. On *CIFAR-100*, similar to previous results of *ResNet-18*, our IAD gains consistent improvements in model robustness.

5.2 Evaluation on *Tiny-ImageNet* Dataset

Experiment Setup. In this part, we evaluate these methods on a more challenging *Tiny-ImageNet* dataset. For these adversarially pre-trained models, we follow the settings of (Chen et al., 2021) to train AT and TRADES. To be specific, we train *PreActive-ResNet-18* using SGD with 0.9 momentum for 100 epochs. The initial learning rate is 0.1 divided by 10 at Epoch 50 and 80 respectively, and the

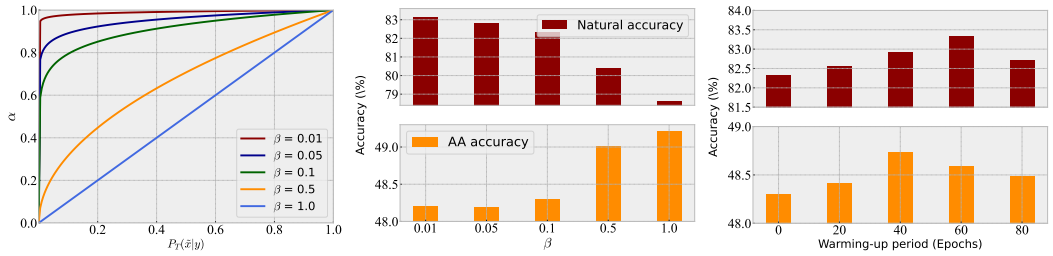


Figure 4: Analysis about using different β and warming-up periods in IAD: *Left*, the values of α (Eq. 4) under different β; *Middle*, Natural and AA accuracy of teacher model using different β; *Right*, Natural and AA accuracy of the teacher model with different warming-up periods.

weight decay=0.0005. For distillation baselines, we keep most settings the same as Section 5.1. For IAD, we use $\lambda = 0.9$, $\beta = 0.1$ and 10 epochs as the warming-up period.

Results. We report the results in Table 4. Overall, our method (IAD or AKD^2+IAD) can still achieve better model robustness than other methods. On *Tiny-ImageNet*, as the dataset is more challenging, the adversarially pre-trained teacher models have low standard and robust accuracy. In this case, the student model might be threatened by a large amount of unreliable guidance. As a result, ARD gets poor performance across these evaluation metrics, since it fully trusts the teacher model. About AKD^2, although it does not sacrifice much natural accuracy, the robust performance is worse than IAD. As for IAD with AKD^2, the results verify that the proposed methods can improve the robustness with no sacrifice on the natural performance in the more challenging dataset.

5.3 Ablation Studies

Experiment Setup. To understand the effects of different β and different warming-up periods on CIFAR-10 dataset, we conduct the ablation study in this part. As for other component of IAD, we may defer to Appendix A for further information. Here, we choose the ResNet-18 as the backbone model, and keep the experimental settings the same as Section 5.1. In the first experiments, we set no warming-up period and study different β. Then, in the second experiments, we set $\beta = 0.1$ and use different warming-up periods.

Results. We report part of the results in Figure 4. The complete results with other evaluation metrics like FGSM, PGD-20 and CW∞ accuracy is put in Appendix A.2. In Figure 4, we first visualize the values of the α using different β in the left panel, which shows the proportion of the teacher guidance and student introspection in adversarial distillation. The bigger the beta corresponds to a larger proportion of the student introspection. In the middle panel, we plot the natural and AA accuracy of the student models distilled by different β. We note that the AA accuracy is improved when the student model trusts itself more with the larger β value. However, the natural accuracy is decreasing along with the increasing of the β value. Similarly, we adjust the length of warming-up periods and check the natural and AA accuracy in the right panel of Figure 4. We find that setting the student model partially trust itself at the beginning of the training process leads to inadequate robustness improvements and more sacrifice on natural accuracy. An appropriate warming-up period at the early stage can improve the student model performance on the adversarial examples.

6 Conclusion

In this paper, we study distillation from adversarially pre-trained models. We take a closer look at adversarial distillation and discover that the guidance of teacher model is progressively unreliable by considering the robustness. Hence, we explore the construction of reliable guidance in adversarial distillation and propose a method for distillation from unreliable teacher models, i.e., Introspective Adversarial Distillation. Our methods encourages the student model partially instead of fully trust the guidance of the teacher model and gradually trust its self-introspection more to improve robustness.

7 ETHICS STATEMENT

This paper does not raise any ethics concerns. This study does not involve any human subjects, practices to data set releases, potentially harmful insights, methodologies and applications, potential conflicts of interest and sponsorship, discrimination/bias/fairness concerns, privacy and security issues, legal compliance, and research integrity issues.

8 REPRODUCIBILITY STATEMENT

To ensure the reproducibility of experimental results, we will provide a link for an anonymous repository about the source codes of this paper in the discussion forums.

REFERENCES

Jean-Baptiste Alayrac, Jonathan Uesato, Po-Sen Huang, Alhussein Fawzi, Robert Stanforth, and Pushmeet Kohli. Are labels required for improving adversarial robustness? In *NeurIPS*, 2019.

Yang Bai, Yuyuan Zeng, Yong Jiang, Shu-Tao Xia, Xingjun Ma, and Yisen Wang. Improving adversarial robustness via channel-wise activation suppressing. In *ICLR*, 2021.

Qi-Zhi Cai, Chang Liu, and Dawn Song. Curriculum adversarial training. In *IJCAI*, 2018.

Nicholas Carlini and David A. Wagner. Towards evaluating the robustness of neural networks. In *Symposium on Security and Privacy (SP)*, 2017.

Tianlong Chen, Sijia Liu, Shiyu Chang, Yu Cheng, Lisa Amini, and Zhangyang Wang. Adversarial robustness: From self-supervised pre-training to fine-tuning. In *CVPR*, 2020.

Tianlong Chen, Zhenyu Zhang, Sijia Liu, Shiyu Chang, and Zhangyang Wang. Robust overfitting may be mitigated by properly learned smoothening. In *ICLR*, 2021.

Mark Craven and Jude Shavlik. Extracting tree-structured representations of trained networks. *NeurIPS*, 1996.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, 2019.

Gavin Weiguang Ding, Yash Sharma, Kry Yik Chau Lui, and Ruitong Huang. Mma training: Direct input space margin maximization through adversarial training. In *ICLR*, 2020.

Xuefeng Du, Jingfeng Zhang, Bo Han, Tongliang Liu, Yu Rong, Gang Niu, Junzhou Huang, and Masashi Sugiyama. Learning diverse-structured networks for adversarial robustness. In *ICML*, 2021.

Micah Goldblum, Liam Fowl, Soheil Feizi, and Tom Goldstein. Adversarially robust distillation. In *AAAI*, 2020.

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

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. In *arXiv*, 2015.

Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Adversarial examples are not bugs, they are features. In *NeurIPS*, 2019.

Ziyu Jiang, Tianlong Chen, Ting Chen, and Zhangyang Wang. Robust pre-training by adversarial contrastive learning. In *NeurIPS*, 2020.

Alex Krizhevsky. Learning multiple layers of features from tiny images. In *arXiv*, 2009.

Ram Shankar Siva Kumar, Magnus Nyström, John Lambert, Andrew Marshall, Mario Goertzel, Andi Comissoneru, Matt Swann, and Sharon Xia. Adversarial machine learning-industry perspectives. In *2020 IEEE Security and Privacy Workshops (SPW)*, 2020.

Ya Le and X. Yang. Tiny imagenet visual recognition challenge. 2015.

Todd Litman. *Autonomous vehicle implementation predictions*. Victoria Transport Policy Institute Victoria, Canada, 2017.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *ICLR*, 2018.

Nicolas Papernot, Patrick McDaniel, Xi Wu, Somesh Jha, and Ananthram Swami. Distillation as a defense to adversarial perturbations against deep neural networks. In *2016 IEEE symposium on security and privacy (SP)*, 2016.

Leslie Rice, Eric Wong, and J Zico Kolter. Overfitting in adversarially robust deep learning. In *ICML*, 2020.

Hadi Salman, Andrew Ilyas, Logan Engstrom, Ashish Kapoor, and Aleksander Madry. Do adversarially robust imagenet models transfer better? In *NeurIPS*, 2020.

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

Qi Tian, Kun Kuang, Kelu Jiang, Fei Wu, and Yisen Wang. Analysis and applications of class-wise robustness in adversarial training. In *KDD*, 2021.

Haotao Wang, Tianlong Chen, Shupeng Gui, Ting-Kuei Hu, Ji Liu, and Zhangyang Wang. Once-for-all adversarial training: In-situ tradeoff between robustness and accuracy for free. 2020a.

Yisen Wang, Xingjun Ma, James Bailey, Jinfeng Yi, Bowen Zhou, and Quanquan Gu. On the convergence and robustness of adversarial training. In *ICML*, 2019.

Yisen Wang, Difan Zou, Jinfeng Yi, James Bailey, Xingjun Ma, and Quanquan Gu. Improving adversarial robustness requires revisiting misclassified examples. In *ICLR*, 2020b.

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

Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *arXiv:1605.07146*, 2016.

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*, 2019.

Jingfeng Zhang, Xilie Xu, Bo Han, Gang Niu, Lizhen Cui, Masashi Sugiyama, and Mohan Kankanhalli. Attacks which do not kill training make adversarial learning stronger. In *ICML*, 2020.

Jingfeng Zhang, Jianing Zhu, Gang Niu, Bo Han, Masashi Sugiyama, and Mohan Kankanhalli. Geometry-aware instance-reweighted adversarial training. In *ICLR*, 2021.

Algorithm 2 AKD^2+IAD

Input: student model S, teacher model T, training dataset $D = \{(x_i, y_i)\}_{i=1}^{n}$, learning rate η, number of epochs N, batch size m, number of batches M, temperature parameter τ, teacher model's predicted probability α, adjustable parameter λ and β.
Output: adversarially robust model S_r
for epoch $= 1, \ldots, N$ **do**
  for mini-batch $= 1, \ldots, M$ **do**
    Sample a mini-batch $\{(x_i, y_i)\}_{i=1}^{m}$ from D
    for $i = 1, \ldots, m$ (in parallel) **do**
      Obtain adversarial data $\tilde{x}_i$ of x_i by PGD based on Eq. equation 2.
      Compute α for each adversarial data based on Eq. equation 4.
    end for

$$\theta \leftarrow \theta - \eta \nabla_\theta \left\{ \begin{array}{l} \lambda_1 \ell_{CE}(S(\tilde{x}), y) + \lambda_2 \cdot \tau^2 \cdot \\ (\alpha \underbrace{\ell_{KL}(S(\tilde{x}_i|\tau)||T(x_i|\tau))}_{\textbf{Teacher Guidance}} + (1-\alpha) \underbrace{\ell_{KL}(S(\tilde{x}_i|\tau)||S(x|\tau))}_{\textbf{Student Introspection}}) + \\ \qquad \lambda_3 \tau^2 \ell_{KL}(S(\tilde{x}|\tau)||T_{st}(\tilde{x}|\tau)) \end{array} \right\}$$

  end for
end for

A Experiment

A.1 Combined algorithm of AKD^2+IAD

Here, we provide the algorithm of AKD^2+IAD used in our experiments.

A.2 Complete results of ablation studies.

Table 5: Test accuracy (%) of IAD using different β.

β	Natural	FGSM	PGD-20	CW$_\infty$	AA
0.01	**83.16%**	**63.97%**	51.28%	50.38%	48.20%
0.05	82.82%	64.12%	51.51%	50.43%	48.19%
0.1	82.32%	63.66%	51.70%	50.51%	48.30%
0.5	80.39%	63.17%	52.62%	50.76%	49.01%
1.0	78.61%	61.96%	**52.81%**	**50.79%**	**49.21%**

Table 6: Test accuracy (%) of IAD using different warming-up periods.

warming-up period	Natural	FGSM	PGD-20	CW$_\infty$	AA
0 epochs	82.32%	63.66%	51.70%	50.51%	48.30%
20 epochs	82.56%	63.68%	51.50%	50.50%	48.41%
40 epochs	82.91%	63.11%	**52.15%**	**50.92%**	**48.73%**
60 epochs	**83.33%**	**63.90%**	51.77%	50.63%	48.59%
80 epochs	82.72%	63.70%	51.75%	50.62%	48.49%

In this part, we report the complete results of our ablation studies in Tables 5 (about β) and 6 (about warming-up periods). In Table 5, we can see that the natural and FGSM accuracy will decrease, and the robust accuracy (PGD-20, CW∞, AA) will increase with the rise of β. In Table 6, we adjust the length of warming-up periods. We can see that letting the student network partially trust itself at the beginning of the training process would result in inadequate robustness improvements.