Adversarial Distillation Based on Slack Matching and Attribution Region Alignment

Shenglin Yin[1], Zhen Xiao[1*], Mingxuan Song[1], Jieyi Long[2]
[1]School of Computer Science, Peking University
[2]Theta Labs, Inc.
{yinsl, songmingxuan}@stu.pku.edu.cn xiaozhen@pku.edu.cn jieyi@thetalabs.org

Abstract

Adversarial distillation (AD) is a highly effective method for enhancing the robustness of small models. Contrary to expectations, a high-performing teacher model does not always result in a more robust student model. This is due to two main reasons. First, when there are significant differences in predictions between the teacher model and the student model, exact matching of predicted values using KL divergence interferes with training, leading to poor performance of existing methods. Second, matching solely based on the output prevents the student model from fully understanding the behavior of the teacher model. To address these challenges, this paper proposes a novel AD method named SmaraAD. During the training process, we facilitate the student model in better understanding the teacher model's behavior by aligning the attribution region that the student model focuses on with that of the teacher model. Concurrently, we relax the condition of exact matching in KL divergence and replace it with a more flexible matching criterion, thereby enhancing the model's robustness. Extensive experiments substantiate the effectiveness of our method in improving the robustness of small models, outperforming previous SOTA methods.

1. Introduction

Deep Neural Networks (DNNs) have achieved significant success in both academic and industrial applications, including image classification [8], face recognition [18], time series forecast [23] and resource scheduling [24, 25]. In pursuit of enhanced performance, present-day deep learning models are frequently designed to be increasingly deep and wide [22]. However, constraints in computational and memory resources pose a challenge to the deployment of these large models, particularly in real-time applications where there is a demand for deploying lightweight models in resource-limited mobile devices for swift inference results. Given the budget limitations inherent in edge deployment, smaller models often lack sufficient protective mechanisms. This deficiency renders them more susceptible to potential threats such as well-orchestrated adversarial attacks [7, 12, 17] for malevolent purposes, in comparison to large models. Consequently, bolstering the robustness of small models against malicious attacks is of critical importance when integrating them into practical applications.

Among the myriad of existing defensive strategies, adversarial training (AT) has been substantiated as one of the most effective approaches [7, 12, 13, 19, 27], garnering significant attention from the research community. Despite its proven reliability in promoting model robustness, AT approaches are not devoid of limitations. Numerous studies illustrate that AT is more proficient with high-volume over-parameterised models as opposed to smaller models [14, 21, 28], implying a direct correlation between model size and robustness. Recently, adversarial distillation (AD) has been put forth as an alternative to augment the robustness of smaller models [6, 32, 33]. Analogous to AT, AD can be framed as a min-max optimisation problem. Its objective is to ensure the student model inherits not only the predictive accuracy but also the adversarial robustness of the robust teacher model within the robust optimisation framework. A summary of the approaches utilised in various state-of-the-art AD approaches is presented in Table 1. Ideally, higher performing teachers should impart more knowledge to their students. However, several studies [3, 29, 33] have observed a lack of direct correlation between the performance of teachers and students, with high-performing teachers not necessarily producing better performing students, and in some cases, even contributing to a decline in student model robustness. *We ascribe this phenomenon to the approach of knowledge transfer between teachers and students.*

Firstly, the conventional approach that employs Kullback-Leibler (KL) divergence to match predictions precisely proves counterproductive when substantial dis-

* Corresponding author.

crepancies exist between the predictions of teacher and student models. This training interference hinders the effectiveness of existing AD approaches, making it difficult for the student model to "understand" the higher-order semantics extracted by the teacher model and limiting the ability to improve the robustness of the student model. Secondly, the prevalent methodology, which matches solely based on output values, fails to enable the student model to apprehend the intricate behaviours exhibited by the teacher model. This results in an incomplete understanding of the pivotal decision-making processes. Consequently, student models are left unable to decipher key nuances and insights from teacher models, thereby obstructing their capacity to learn robust representations.

In response to these challenges and with the aim to bolster the robustness of smaller models, this paper introduces an innovative AD approach (SmaraAD). Our strategy seeks to offer student models a more profound comprehension of the teacher model's behaviour, thereby facilitating superior knowledge transfer. To achieve this, we align the attribution region of the student model, that is, the region of interest the student model concentrates on, with the teacher model's attribution region during the training phase. Consequently, the student model can focus on the same pertinent features as the teacher model and glean valuable information, hence gaining a more thorough understanding of the teacher model's decision-making process. Furthermore, we relax the precise matching condition in the KL divergence, preferring a more lenient matching criterion. In knowledge transfer from teachers to students, our central concern is the correlation degree of predicted outcomes between the two. Therefore, we adopt the Spearman correlation coefficient as a novel matching approach, substituting the KL divergence. This not only mitigates training interference due to variations in predicted values but also fortifies the model's resilience against uncertainty and noise. To corroborate the effectiveness of our suggested approach, we executed comprehensive experiments. The experimental outcomes demonstrate that our AD approach holds considerable advantages in enhancing the robustness of smaller models, outperforming previous state-of-the-art approaches. These empirical results offer robust evidence for the potential applicability of our proposed novel approach in real-world scenarios.

In summary, our main contributions are:

- In an effort to tackle the identified issues and enhance the robustness of smaller models, we introduce an innovative AD technique. This approach seeks to foster superior knowledge transfer by imparting the student model with a deeper insight into the teacher model's behavior. During the training, we strive to accomplish this aim by synchronizing the attribution region of the student model with that of the teacher model. Consequently, the student model can focus on the same pertinent features as the teacher model and accrue a more thorough understanding of the teacher model's decision-making mechanism.
- We employ the Spearman correlation coefficient as a fresh matching substitute for KL divergence. This aids in alleviating training interference attributed to discrepancies in predicted values and bolstering model resilience against uncertainty and noise.
- We provide empirical evidence to substantiate the effectiveness of our suggested approach in improving the performance of smaller models. A multitude of experimental results affirm that our approach significantly surpasses state-of-the-art AT and AD approaches across various scenarios.

2. Related Work

2.1. Adversarial Attack

Adversarial attacks are a form of attack on artificial intelligence models achieved by modifying the input data, leading to incorrect output from the model. Following the introduction of the concept of adversarial examples by Sezgedy et al. [17], numerous studies have investigated model robustness and presented effective methods for adversarial attacks. The Fast Gradient Sign Method (FGSM) [7] is a classical method for adversarial attacks that leverages the model's gradient to generate adversarial examples. Madry et al. [12] proposed Projected Gradient Descent (PGD), a multi-step variant of FGSM. Carlini Wagner et al. [2] introduced an optimization-based method that enables broad usage in assessing the robustness of deep learning models. Croce et al. [5] introduced two enhanced PGD attack methods: APGD-CE and APGD-DLR. Subsequently, they integrated these methods with two complementary black-box attack methods (FAB [4] and Square [1]) to assess their robustness, known as AutoAttack (AA). AA can be considered the most potent attack currently available.

2.2. Adversarial Distillation

Knowledge distillation (KD) is a widely recognized technique for compressing deep neural networks, which has received significant attention in recent years. It is valued for its capability to transfer superior model performance to other models. By utilizing a trained teacher T, KD trains a student S by solving the following optimisation problem:

$$\underset{\theta_S}{argmin}(1-\alpha)\mathcal{L}(S(x),y)+\alpha\tau^2 KL(S^\tau(x),T^\tau(x)), \quad (1)$$

where KL is the Kullback-Leibler divergence and τ is the temperature constant. In contrast to traditional KD, AD places emphasis on the transfer of both clean accuracy and the robustness of teacher models to student models. Goldblum et al. [6] conducted an investigation into the transfer of adversarial robustness from teachers to students in

Table 1. The optimization process of the standard adversarial training method, five types of adversarial distillation methods, and our proposed method. $\mathcal{L}_{min}$ represents the loss function for outer minimization, while $\mathcal{L}_{max}$ represents the loss function for inner maximization. S and T denote the student and teacher networks, respectively. α is a hyperparameter that balance the losses. L_{sm} and L_{align} are the slack matching and attribution region alignment methods we introduced.

Method	$\mathcal{L}_{min}$	$\mathcal{L}_{max}$
Standard-AT	$CE(f(x'), y)$	$CE(f(x'), y)$
ARD	$(1-\alpha)CE(S^\tau(x), y) + \alpha\tau^2 KL(S^\tau(x'), T^\tau(x))$	$CE(S(x'), y)$
IAD	$T_y(x')^\alpha KL(S^\tau(x'), T^\tau(x)) + (1 - T_y(x')^\alpha) KL(S^\tau(x'), S^\tau(x))$	$CE(S(x'), y)$
RSLAD	$(1-\alpha)KL(S(x), T(x)) + \alpha KL(S(x'), T(x))$	$KL(S(x'), T(x))$
MTARD	$(1-\alpha)KL(S(x), T_{nat}(x)) + \alpha KL(S(x'), T_{adv}(x'))$	$CE(S(x'), y)$
AdaAD	$(1-\alpha)KL(S(x), T(x)) + \alpha KL(S(x'), T(x'))$	$KL(S(x'), T(x'))$
SmaraAD (Proposed)	$(1-\alpha)\cdot(L_{sm}(x) + L_{align}(x)) + \alpha\cdot(L_{sm}(x') + L_{align}(x'))$	$KL(S(x'), T(x')) + L_{align}(x')$

the context of KD. They introduced the Adversarial Robust Distillation (ARD) method to imbue student networks with robustness. Zi et al. [33] presented a novel method for distilling adversarial robustness called Robust Soft-Label Adversarial Distillation (RSLAD). RSLAD utilizes robust soft labels, generated by large teacher models trained adversarially, to guide students in learning both natural and adversarial examples across various loss conditions. Zhu et al. [32] asserted that teacher models may not always be consistently reliable and subsequently introduced Introspective Adversarial Distillation (IAD) as a means of achieving reliable AD. Zhao et al. [30] introduced Multi-Teacher Adversarial Robustness Distillation (MTARD) to guide smaller models in adversarial scenarios. Huang et al. [9] proposed Adaptive Adversarial Distillation (AdaAD), where the teacher model engages in an interactive knowledge optimization process with the student model to adaptively search for internal outcomes.

3. The Proposed Approach

In this section, we revisit state-of-the-art AD methods and analyse the limitations of existing AD methods through empirical experiments. We then introduce AD methods based on **S**lack **M**atching and **A**ttribution **R**egion **A**lignment (SmaraAD).

3.1. Limitations of Existing AD Methods

We present a comprehensive summary of various existing AD methods. Table 1 provides a detailed overview of the AD methods we investigated, which generally employ exact matching of hard or soft labels for knowledge transfer between teacher and student models. However, we identified certain limitations in the current AD approaches, particularly concerning the robustness of the student model. To investigate these limitations thoroughly, we conducted a series of empirical experiments.

First, we utilized teacher models with varying degrees of robustness for different AD methods. Then, we compared the robustness of student models trained using different AD methods with that of the corresponding teacher models. Surprisingly, the experimental results depicted in Figure 1 reveal that the robustness of the student model does not improve with an increase in the teacher model's robustness; instead, the robustness of the student model may even decline in certain instances. This observation indicates that existing AD methods still face challenges in achieving robustness for the student model. KL divergence is an exact matching method that prioritizes replicating the teacher model's predictions precisely, rather than capturing its intrinsic patterns. When there are large differences in the predictions between the teacher model and the student model, the student model, in its attempt to precisely match the predictions of the teacher model, may overfit the noise of the teacher model at the expense of the real goal-understanding and learning from the teacher model. This explains why employing KL divergence reduces the model's robustness when there are considerable prediction disparities between the teacher model and the student model.

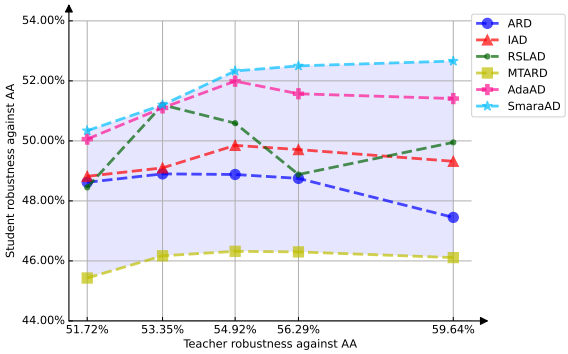

Figure 1. Robustness of AA attacks on ResNet-18 students trained by different AD methods using 5 different teachers. The experiment was done on the CIFAR-10 dataset.

Second, to delve deeper into the disparity between the knowledge acquired by the student model and the teacher model, we employed the class activation mapping (CAM) [31] method. The experimental results, shown in Figure 2, indicate a notable difference between the student and teacher models when comparing the attribution regions.

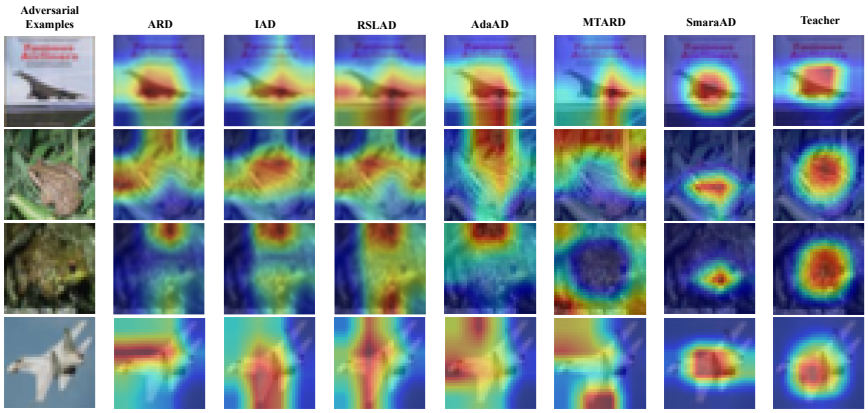

Figure 2. Attribution regions of student models trained by different AD methods and teacher model on adversarial examples.

This suggests that the student model incompletely captures the behavior of the teacher model. Neural networks tightly couple feature extraction and decision-making. Specific input features significantly impact the model's decision outcome, often termed the decision's attribution region. We hypothesise that if two models (teacher model and student model) focus on different feature regions when processing the same inputs, their behavioural properties and decision logics may differ even if their predictions are similar. Existing AD methods prioritize ensuring the consistency between the predictions of the teacher model and the student model, overlooking potential behavioral differences. Consequently, the student model may not fully mimic the behavior of the teacher model, particularly when faced with new, previously unseen data during training.

To address the above issues, we propose two strategies: slack matching with Spearman correlation and behavioural learning with attribution region alignment.

3.2. Slack Matching with Spearman Correlation

Empirical experiments have demonstrated a counterintuitive phenomenon: when using KL divergence as the approach, the performance of the student model declines as the performance of the teacher model improves. This may be attributed to the requirement of exact matching through KL-based methods, which compels the student model to over-approximate the predicted probability distribution of the teacher model. Inspired by Huang et al. [10], it is crucial to prioritize maintaining predictive relationships between teacher and student models, especially the relative prediction rankings, while transferring knowledge from teachers to students. Accordingly, we employed the Spearman correlation coefficient [16] in place of the KL divergence. This selection permits a more flexible matching relationship between the teacher and student models, facilitating the assessment of correlation between their predictions.

The Spearman correlation coefficient is a measure of correlation between two variables. In other words, when one variable changes, the value of the other variable tends to change in the same direction, without highlighting the consistency of its absolute value. This feature mitigates the issue of over-approximating the teacher model during training, reducing training interference and facilitating improved learning of the student model from the teacher model. By maximizing the Spearman correlation coefficient, we are able to effectively transfer the knowledge of the teacher model to the student model. The Spearman correlation coefficient ρ between two random variables $\mathbf{u}$ and $\mathbf{v}$ is calculated as follows:

$$\rho(\mathbf{u}, \mathbf{v}) = 1 - \frac{6\sum_{i=0}^{n-1} d_i^2}{n(n^2-1)}, \tag{2}$$

$$d_i = rank(u_i) - rank(v_i), \tag{3}$$

where n is the number of classes. d_i is the difference in the ranking of the variables, i.e., the difference in the ranking of the two variables at each data point. $rank(u_i)$ denotes the ranking of the i-th data point in $\mathbf{u}$.

It is important to note that $rank(\cdot)$ uses hard ranking (i.e., sorting the data values and giving them an exact rank), an operation that is not differentiable and can affect the process of model convergence. To address this, we utilize a continuous function to approximate the ranking function, making the entire computation differentiable. For u_i, its soft ranking can be defined as follows:

$$soft\_rank(u_i) = \sigma(\gamma(\mathbf{u} - u_i)), \tag{4}$$

where $\sigma(\cdot)$ is Sigmoid function. γ is a constant indicating the degree of "softening". As $\gamma \to \infty$, the soft ranking approaches the hard ranking. Utilizing this soft ranking definition, we can employ it in Spearman's original equation, supplanting the hard ranking $rank(\cdot)$.

In the slack matching process, we want to increase the positive correlation between the student model and the teacher model, so we can directly transfer knowledge from

the teacher model to the student model by maximizing the Spearman correlation coefficient between the two. Specifically, assuming that $T(x)$ and $S(x)$ denote the predicted probability distributions of the teacher model and the student model for the input sample x, respectively, then our optimisation objective is as follows:

$$L_{sm}(x) = -\rho(S(x), T(x)). \tag{5}$$

3.3. Behavioral Learning with Attribution Region Alignment

In order to achieve the understanding of the student model to the teacher model's behavior, we use CAM to compute the attribution regions between the two models. CAM is an interpretable technique for image classification tasks that visualises the neural network's regions of attention on an image to understand the regions that the network is focusing on when making predictions. For each input image, CAM generates a heat map representing the importance of different regions. To compute the attribution regions, we consider a given input x and utilize $A_T^k(x)$ and $A_S^k(x)$ as the k-th feature maps of the last layer for the teacher model and student model, respectively. The computation of attribution regions involves the following steps:

$$\begin{aligned} CAM_T(x) &= \sum_k w_T^k A_T^k(x), \\ CAM_S(x) &= \sum_k w_S^k A_S^k(x), \end{aligned} \tag{6}$$

where w_T^k and w_S^k are the weights of the corresponding feature maps for the teacher model and student model, respectively. These weights come from the classification layer and indicate the importance of the model for each feature map.

In the training phase, we achieved attribution region alignment by minimising the Mean Squared Error (MSE) between the attribution regions of the two models, thus facilitating the imitation of the teacher model's behaviour by the student model, which was calculated as follows:

$$\begin{aligned} L_{align}(x) = \frac{1}{W \times H} \sum_{i=1}^{W} \sum_{j=1}^{H} \\ (G(CAM_T(x))_{ij} - G(CAM_S(x))_{ij})^2, \end{aligned} \tag{7}$$

where W and H are width and height of the attribution regions, respectively. Please note that the width and height of the attribution region may not be the same for the student model and the teacher model because of their different structures. Hence, to address this disparity, we utilize a transformation method $G(\cdot)$ to equalize the width and height of $CAM_T(x)$ and $CAM_S(x)$. Specifically, we employ bilinear interpolation as the transformation method $G(\cdot)$. It can be easily implemented using $F.interpolate()$ in Pytorch.

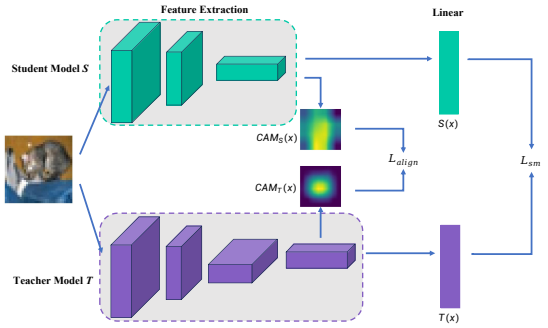

Figure 3. The process of outer optimization.

3.4. Optimization Process

The optimization process involves two crucial steps: inner maximization and outer minimization, facilitating efficient knowledge transfer.

Inner maximization is utilized to generate adversarial examples. These examples aim to represent the most significant differences between the student and teacher models, thereby providing deeper insights into the student model's weaknesses in various scenarios and ultimately enhancing its robustness. Therefore, we maximise the difference between the KL loss and the attribution region between the two. The equation for generating the adversarial example is as follows:

$$x' = \underset{\|x'-x\|_p \leq \epsilon}{argmax} \, (L_{KL} + L_{align}). \tag{8}$$

Outer minimization optimises the student model and aims to transfer knowledge from the teacher model to the student model. In this process, we employ the aforementioned methods of slack matching and attribution region alignment as loss functions to optimize the student model. Our objective during optimization is to maximize the linear correlation and attribution region similarity between the student and teacher model outputs. The outer optimization process is depicted in Figure 3, and the minimization equation is presented below:

$$\begin{aligned} L_{sm\&align} = & c + (1-\alpha) \cdot (L_{sm}(x) + L_{align}(x)) + \\ & \alpha \cdot (L_{sm}(x') + L_{align}(x')), \end{aligned} \tag{9}$$

where $c \geq 1$ is a constant, x' is the adversarial example generated by inner maximization and $\alpha \in (0, 1)$ is a hyperparameter.

The reliability of the distillation process is compromised by the consideration that the teacher model may not consistently make accurate predictions for certain inputs. Studies [9, 32] have shown that the teacher model's reliability declines during distillation training, which encourages the student model to adopt a cautious reliance on the teacher. Following the approach of [32], we classify two scenarios based on whether the teacher model can correctly classify

adversarial examples generated based on the student model: if the teacher model predicts correctly, we assign weights to each sample proportional to the teacher model's confidence levels. Conversely, if the teacher model predicts incorrectly, the student model is directed towards increased self-learning. The equation for outer optimization is presented below:

$$L_{outer} = \underbrace{(P_T(x'|y))^{\lambda} \cdot w \cdot L_{sm\&align}}_{\text{Teacher Guidance}} + \underbrace{(1-(P_T(x'|y))^{\lambda}) \cdot \left(\rho\left(S(x)||S(x')\right) + CE(S(x),y)\right)}_{\text{Student Self-learning}}, \tag{10}$$

where $P_T(\cdot|y)$ is the teacher model's prediction probability for the target label y, and $\lambda \in (0,1)$ is a hyperparameter that sharpens the prediction. When the teacher model's prediction $(P_T(x'|y))$ for the true label y of the adversarial example x' is low, the weight of Teacher Guidance is reduced. w represents the weight of the input.

Within this framework, we consider prediction uncertainty from the perspective of information entropy and allocate weights to each sample accordingly. For a classification task, the model's predicted probability distribution over categories is $P = [p_1, p_2, \ldots, p_C]$, and the prediction uncertainty is quantified by the information entropy H, defined as $H(P) = -\sum_{i=1}^{C} p_i \log(p_i)$. The entropy reaches its maximum value $H_{\max} = \log(C)$ when the predicted probabilities are equal across all categories. By calculating $H_{\text{norm}} = \frac{H}{H_{\max}}$, we normalize the entropy values relative to the maximum possible value given the number of categories. The setting of the sample weight w is to map the inverse relationship of entropy within a predefined weight range $(0,1)$. The calculation is $w = 1 - H_{\text{norm}}$ and is constrained by $w = \text{clamp}(w, 0, 1)$. This allows for lower weights on samples with higher uncertainty, aiding the model's learning from difficult samples.

4. Experiments

4.1. Experimental Setup

We evaluated the effectiveness of SmaraAD using two benchmark image datasets: CIFAR-10 and CIFAR-100 [11]. We conducted a comparison between proposed method and three adversarial training (AT) methods, namely PGD-AT, TRADES and LAS-AT, as well as several representative AD methods, including ARD, IAD, RSLAD, MTARD, and AdaAD. In terms of model selection, adhering to the standard setup of AD, we considered two student models: ResNet-18 [8] and MobileNet-V2 [15]. As for the teacher models, we selected adversarially trained teacher models: WideResNet-28-10 [26] and WideResNet-34-10 for CIFAR-10 and WideResNet-34-10 for CIFAR-100. In training the teacher model, we used the additional dataset generated by Wang et al. [20] The performance of the teacher models is shown in Table 2.

4.2. Implementation Details

SmaraAD (employing Eq. 9 for outer optimization) is our approach without considering the accuracy of the teacher model's predictions, while SmaraAD++ (employing Eq. 10 for outer optimization) is our approach with considering the accuracy of the teacher model's predictions. We employ an SGD optimizer with momentum for training the model, setting the initial learning rate to 0.1, momentum to 0.9, and weight decay to 5e-4. The batch size is set to 256. For AT and other AD methods, the training process comprises 200 epochs, and we decrease the learning rate by a factor of 10 at the 115th, 160th and 185th epoch. During the internal optimization, we run 10 epochs with a step size of 2/255 and a total perturbation limit of 8/255 under L_∞ constraints. In our method, the hyperparameter α is set to 1.0 in SmaraAD, as suggested by [6, 9], and similarly, λ is also assigned the value of 0.1, following the recommendation in [32]. The constant c is set to 2 and γ is set to 1000.

4.3. Evaluation Metrics

After completing the training process, we assess the model's performance by measuring its accuracy on the natural samples (referred to as clean accuracy) as well as its resilience to adversarial attacks on the adversarial examples (referred to as robust accuracy). We selected multiple adversarial attack methods to evaluate the trained model's robustness. These methods include FGSM, PGD, C&W$_\infty$ (optimized by PGD), and AutoAttack (AA). It is important to note that these attacks represent commonly used adversarial attack methods for evaluating the robustness of models in the field of adversarial robustness. Regarding FGSM, PGD, C&W$_\infty$, and AA, we set the maximum perturbation size to 8/255. Furthermore, we employ 20 steps for PGD and C&W$_\infty$, each with a step size of 2/255.

4.4. Adversarial Robustness Evaluation

4.4.1 White-box Robustness.

Tables 3 and 4 present the white-box robustness results of our proposed methods compared to other methods on the CIFAR-10 and CIFAR-100 datasets, respectively. In the CIFAR-10, taking the combination of the ResNet-18 student model and the WideResNet-34-10 teacher model as an example, SmaraAD's accuracy in the FGSM attack scenario is 77.56%, which is 2.52% higher than AdaAD's 75.04%, and in the PGD-20 attack, its accuracy is 57.47%, surpassing LAS-AT's 53.98% by 3.49%. Under the combination of the MN-V2 student model and the WideResNet-28-10 teacher model, SmaraAD++ achieves an accuracy of 54.07% in the C&W$_\infty$ attack, which is 1.85% higher than

Table 2. The performance of teacher models for two datasets.

Dataset	Teacher	Clean	FGSM	PGD-20	C&W$_\infty$	AA
CIFAR-10	WRN-28-10	86.89%	79.42%	57.82%	56.73%	55.34%
CIFAR-10	WRN-34-10	88.15%	80.42%	63.51%	60.32%	59.87%
CIFAR-100	WRN-34-10	66.41%	52.78%	38.12%	37.76%	35.06%

Table 3. White-box robustness results on CIFAR-10 dataset. The maximum adversarial perturbation is $\epsilon = 8/255$. The best results are **boldfaced**, and the second best results are underlined.

Teacher Model		WideResNet-34-10					WideResNet-28-10				
Model	Method	Clean	FGSM	PGD-20	C&W$_\infty$	AA	Clean	FGSM	PGD-20	C&W$_\infty$	AA
RN-18	PGD-AT	83.32%	56.67%	49.79%	48.61%	46.90%	83.32%	56.67%	49.79%	48.61%	46.90%
	TRADES	82.96%	57.69%	52.34%	50.12%	49.20%	82.96%	57.69%	52.34%	50.12%	49.20%
	LAS-AT	82.73%	60.21%	53.98%	52.09%	50.22%	82.73%	60.21%	53.98%	52.09%	50.22%
	ARD	81.47%	72.17%	52.53%	50.28%	48.61%	82.51%	71.95%	52.92%	52.43%	49.02%
	IAD	81.49%	72.36%	52.54%	50.04%	49.87%	83.33%	72.67%	52.78%	51.63%	49.30%
	RSLAD	82.81%	72.29%	53.68%	51.31%	49.95%	83.50%	73.30%	53.75%	52.48%	50.49%
	MTARD	86.84%	73.64%	50.87%	47.93%	46.12%	**87.11%**	74.04%	51.63%	48.12%	46.38%
	AdaAD	85.26%	75.04%	55.40%	52.82%	51.41%	85.48%	75.34%	55.45%	53.84%	51.99%
	SmaraAD	86.31%	77.56%	57.47%	55.20%	52.66%	85.77%	76.48%	55.70%	54.43%	52.36%
	SmaraAD++	**86.90%**	**78.18%**	**58.58%**	**56.75%**	**53.40%**	86.70%	**77.47%**	**56.03%**	**55.21%**	**52.88%**
MN-V2	PGD-AT	80.10%	54.51%	49.11%	48.20%	45.67%	80.10%	54.51%	49.11%	48.20%	45.67%
	TRADES	79.98%	55.58%	51.49%	50.40%	46.95%	79.98%	55.58%	51.49%	50.40%	46.95%
	LAS-AT	80.62%	57.74%	52.87%	51.12%	48.46%	80.62%	57.74%	52.87%	51.12%	48.46%
	ARD	82.21%	70.97%	51.95%	49.10%	48.32%	83.43%	71.90%	52.01%	50.58%	49.50%
	IAD	81.98%	69.85%	52.28%	49.02%	47.31%	82.80%	72.55%	52.24%	49.80%	47.94%
	RSLAD	82.10%	71.06%	52.32%	49.36%	48.29%	83.26%	72.93%	52.67%	51.27%	49.41%
	MTARD	**87.43%**	71.33%	42.15%	41.76%	40.43%	**88.86%**	72.99%	43.78%	43.28%	41.99%
	AdaAD	83.42%	72.89%	54.19%	51.72%	49.53%	84.42%	73.88%	54.69%	52.22%	50.03%
	SmaraAD	85.47%	74.58%	56.31%	54.30%	52.17%	85.12%	74.05%	55.29%	53.95%	51.88%
	SmaraAD++	86.66%	**76.24%**	**57.17%**	**55.19%**	**52.54%**	86.48%	**75.36%**	**55.84%**	**54.07%**	**52.00%**

Table 4. White-box robustness results on CIFAR-100 dataset. The maximum adversarial perturbation is $\epsilon = 8/255$. The best results are **boldfaced**, and the second best results are underlined.

Teacher Model		WideResNet-34-10				
Model	Method	Clean	FGSM	PGD-20	C&W$_\infty$	AA
RN-18	PGD-AD	55.47%	40.08%	25.72%	23.98%	21.30%
	TRADES	55.98%	42.58%	26.88%	25.40%	23.95%
	LAS-AT	56.62%	44.19%	27.93%	27.41%	25.33%
	ARD	60.34%	44.71%	30.49%	29.86%	26.03%
	IAD	60.72%	44.85%	30.79%	29.98%	25.91%
	RSLAD	62.32%	46.45%	32.33%	30.30%	27.56%
	MTARD	**63.87%**	45.87%	25.56%	24.55%	22.38%
	AdaAD	61.72%	46.44%	32.89%	30.15%	27.18%
	SmaraAD	63.18%	49.42%	34.13%	33.03%	29.16%
	SmaraAD++	62.64%	**50.82%**	**35.79%**	**34.03%**	**30.63%**
MN-V2	PGD-AD	55.94%	39.78%	25.98%	23.63%	21.18%
	TRADES	55.97%	40.02%	26.00%	24.83%	23.70%
	LAS-AT	56.02%	43.75%	27.35%	27.28%	25.19%
	ARD	60.88%	44.42%	29.69%	29.90%	25.92%
	IAD	60.79%	44.58%	30.43%	28.51%	25.74%
	RSLAD	64.20%	46.91%	31.14%	28.77%	26.62%
	MTARD	**66.01%**	47.18%	26.05%	25.61%	25.18%
	AdaAD	62.30%	47.98%	32.13%	29.89%	27.01%
	SmaraAD	64.83%	48.43%	33.57%	32.92%	28.81%
	SmaraAD++	65.71%	**49.58%**	**35.12%**	**33.89%**	**29.65%**

Table 5. Black-box robustness results on CIFAR-10 dataset. The maximum adversarial perturbation is $\epsilon = 8/255$. The best results are **boldfaced**, and the second best results are underlined.

Robustness of teacher	70.64%		59.04%	
Method	PGD-20	C&W$_\infty$	PGD-20	C&W$_\infty$
LAS-AT	65.96%	65.01%	61.32%	61.09%
RSLAD	67.93%	67.50%	63.62%	63.36%
AdaAD	68.77%	67.80%	64.57%	64.45%
SmaraAD	**69.16%**	**68.29%**	**65.20%**	**65.01%**

AdaAD's 52.22%, and an accuracy of 52.00% in the AA attack, outperforming RSLAD's 49.41% by 2.59%.

On the CIFAR-100, under the RN-18 model, SmaraAD's accuracy in the FGSM attack is 49.42%, which is 2.98% higher than AdaAD's 46.44%, and in the PGD-20 attack, its accuracy is 34.13%, exceeding AdaAD's 32.89% by 1.24%. Under the MN-V2 model, SmaraAD++ achieves an accuracy of 35.12% in the PGD-20 attack and 29.65% in the AA attack, both higher than any other methods.

These results collectively demonstrate that SmaraAD and SmaraAD++ exhibit exceptional robustness against various types of attacks, underscoring their effectiveness in ensuring model security and robustness, particularly in diverse adversarial attack environments.

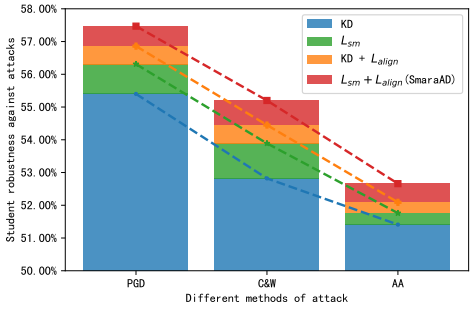

Figure 4. Ablation studies using our SmaraAD and its variant distilled ResNet-18 student model. KD: trained using only KL divergence; KD+L_{align}: trained using both KD and attribution region alignment; L_{sm}: trained using only slack matching; L_{sm}+L_{align}: trained using our SmaraAD.

4.4.2 Black-box Robustness.

We evaluated the black-box robustness of several methods with strong defences. We tested the transfer attack on the CIFAR-10 dataset. Specifically, we use the teacher model with 70.64% (WRN-34-20) and 59.04% (RN-18) robustness on PGD-20 attacks as a proxy model to generate samples against PGD-20 and C&W$_\infty$. Table 5 reports the evaluation results. It can be observed that that when the structure of the agent model is similar to that of the student model, the higher the success rate of the attack. In all black-box attacks, our SmaraAD outperforms all baseline methods, which proves the superiority of our AD method.

4.5. Ablation Study

In order to better understand the impact of each component of the SmaraAD on robustness, we conducted a set of ablation studies using a ResNet-18 student model on the CIFAR-10 dataset. We replaced L_{sm} with KL divergence and tested the robustness of the student model after training. Additionally, we removed the L_{algin} and trained the student model solely using L_{sm}. The results of the ablation study are reported in Figure 4. From the experimental results, it can be observed that the robustness of the student model is best when both L_{algin} and L_{sm} are present. This confirms the importance of each component of the SmaraAD.

4.6. Further Explorations

4.6.1 The Effect of Teacher Model Performance on Student Models.

We conducted experiments to investigate the impact of the teacher on the robustness of the student. This experiment was performed using a ResNet-18 student model on the CIFAR-10 dataset, and we studied the robustness extracted from six different teacher models using various AD methods. The results are shown in Figure 1. From the experimental results, it can be observed that when using KL-based AD methods, the student's robustness decreases as the performance of the teacher model increases. However, our SmaraAD allows the robustness of the distilled student model to increase with the size of the teacher model. This also confirms the effectiveness of our SmaraAD.

4.6.2 Attribution Region Learned by SmaraAD.

We use CAM to visualize the attribution regions of the model, providing an intuitive assessment of the similarity between the knowledge learned by the student and the teacher model. In the same adversarial examples, a higher similarity between the student's attribution regions and those of the teacher indicates a more consistent behavior of the student model with the teacher model. Taking the example of a ResNet-18 student model distilled from a WideResNet-28-10 teacher model on the CIFAR-10 dataset, we visualize the attribution regions as shown in Figure 2. It can be observed that compared to the baseline methods ARD, IAD, RSLAD, MTARD and AdaAD, the attribution regions of the student trained with our SmaraAD are significantly more similar to those of the teacher. This indicates that the student trained with our SmaraAD can indeed better mimic the behavior of the teacher and acquire more knowledge from the teacher.

5. Conclusion

This paper explores how AD can be used to enhance the robustness of small models. It is shown that a high-performing teacher model does not always guarantee a more robust student model. There are two main reasons for this discrepancy. Firstly, existing methods face challenges in dealing with significant predictive differences between teacher and student models. Using KL divergence for exact matching of predictions during training leads to poor performance. Second, relying solely on output-based matching can prevent the student model from fully capturing the behaviour of the teacher model. To address these challenges, we introduce a novel AD method in this study. Our approach helps student models understand the behaviour of teacher models by aligning their respective attribution regions. In addition, we employ a more relaxed matching instead of exact matching in KL divergence, which improves the robustness of the model. Extensive experiments demonstrate the effectiveness of our proposed approach in enhancing the robustness of small models beyond previous SOTA techniques.

6. Acknowledgment

The authors would like to thank the anonymous reviewers for their comments. This work was supported by the Beijing Natural Science Foundation under Funding No. IS23055. The contact author is Zhen Xiao.

References

[1] Maksym Andriushchenko, Francesco Croce, Nicolas Flammarion, and Matthias Hein. Square attack: a query-efficient black-box adversarial attack via random search. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIII*, pages 484–501. Springer, 2020. 2

[2] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 ieee symposium on security and privacy (sp)*, pages 39–57. Ieee, 2017. 2

[3] Defang Chen, Jian-Ping Mei, Hailin Zhang, Can Wang, Yan Feng, and Chun Chen. Knowledge distillation with the reused teacher classifier. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11933–11942, 2022. 1

[4] Francesco Croce and Matthias Hein. Minimally distorted adversarial examples with a fast adaptive boundary attack. In *International Conference on Machine Learning*, pages 2196–2205. PMLR, 2020. 2

[5] Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *International conference on machine learning*, pages 2206–2216. PMLR, 2020. 2

[6] Micah Goldblum, Liam Fowl, Soheil Feizi, and Tom Goldstein. Adversarially robust distillation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3996–4003, 2020. 1, 2, 6

[7] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014. 1, 2

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1, 6

[9] Bo Huang, Mingyang Chen, Yi Wang, Junda Lu, Minhao Cheng, and Wei Wang. Boosting accuracy and robustness of student models via adaptive adversarial distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24668–24677, 2023. 3, 5, 6

[10] Tao Huang, Shan You, Fei Wang, Chen Qian, and Chang Xu. Knowledge distillation from a stronger teacher. *Advances in Neural Information Processing Systems*, 35:33716–33727, 2022. 4

[11] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 6

[12] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017. 1, 2

[13] Pratyush Maini, Eric Wong, and Zico Kolter. Adversarial robustness against the union of multiple perturbation models. In *International Conference on Machine Learning*, pages 6640–6650. PMLR, 2020. 1

[14] Tianyu Pang, Xiao Yang, Yinpeng Dong, Hang Su, and Jun Zhu. Bag of tricks for adversarial training. *arXiv preprint arXiv:2010.00467*, 2020. 1

[15] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018. 6

[16] Charles Spearman. " general intelligence" objectively determined and measured. 1961. 4

[17] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013. 1, 2

[18] Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu. Cosface: Large margin cosine loss for deep face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5265–5274, 2018. 1

[19] Yisen Wang, Difan Zou, Jinfeng Yi, James Bailey, Xingjun Ma, and Quanquan Gu. Improving adversarial robustness requires revisiting misclassified examples. In *International Conference on Learning Representations*, 2020. 1

[20] Zekai Wang, Tianyu Pang, Chao Du, Min Lin, Weiwei Liu, and Shuicheng Yan. Better diffusion models further improve adversarial training. *arXiv preprint arXiv:2302.04638*, 2023. 6

[21] Boxi Wu, Jinghui Chen, Deng Cai, Xiaofei He, and Quanquan Gu. Do wider neural networks really help adversarial robustness? *Advances in Neural Information Processing Systems*, 34:7054–7067, 2021. 1

[22] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1492–1500, 2017. 1

[23] Mingzhe Xing, Shuqing Bian, Wayne Xin Zhao, Zhen Xiao, Xinji Luo, Cunxiang Yin, Jing Cai, and Yancheng He. Learning reliable user representations from volatile and sparse data to accurately predict customer lifetime value. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 3806–3816, 2021. 1

[24] Mingzhe Xing, Hangyu Mao, and Zhen Xiao. Fast and fine-grained autoscaler for streaming jobs with reinforcement learning. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence (Vienna, Austria, 23-29 July 2022)(IJCAI 2022). ijcai. org, USA*, pages 564–570, 2022. 1

[25] Mingzhe Xing, Hangyu Mao, Shenglin Yin, Lichen Pan, Zhengchao Zhang, Zhen Xiao, and Jieyi Long. A dual-agent scheduler for distributed deep learning jobs on public cloud via reinforcement learning. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 2776–2788, 2023. 1

[26] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016. 6

[27] 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 *International conference on machine learning*, pages 11278–11287. PMLR, 2020. 1

[28] Jingfeng Zhang, Jianing Zhu, Gang Niu, Bo Han, Masashi Sugiyama, and Mohan Kankanhalli. Geometry-aware instance-reweighted adversarial training. *arXiv preprint arXiv:2010.01736*, 2020. 1

[29] Borui Zhao, Quan Cui, Renjie Song, Yiyu Qiu, and Jiajun Liang. Decoupled knowledge distillation. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pages 11953–11962, 2022. 1

[30] Shiji Zhao, Jie Yu, Zhenlong Sun, Bo Zhang, and Xingxing Wei. Enhanced accuracy and robustness via multi-teacher adversarial distillation. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part IV*, pages 585–602. Springer, 2022. 3

[31] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016. 3

[32] Jianing Zhu, Jiangchao Yao, Bo Han, Jingfeng Zhang, Tongliang Liu, Gang Niu, Jingren Zhou, Jianliang Xu, and Hongxia Yang. Reliable adversarial distillation with unreliable teachers. *arXiv preprint arXiv:2106.04928*, 2021. 1, 3, 5, 6

[33] Bojia Zi, Shihao Zhao, Xingjun Ma, and Yu-Gang Jiang. Revisiting adversarial robustness distillation: Robust soft labels make student better. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16443–16452, 2021. 1, 3