HOW AND WHEN ADVERSARIAL ROBUSTNESS TRANSFERS IN KNOWLEDGE DISTILLATION?

Anonymous authors
Paper under double-blind review

ABSTRACT

Knowledge distillation (KD) has been widely used in teacher-student training, with applications to model compression in resource-constrained deep learning. Current works mainly focus on preserving the accuracy of the teacher model. However, other important model properties, such as adversarial robustness, can be lost during distillation. This paper studies how and when the adversarial robustness can be transferred from a teacher model to a student model in KD. We show that standard KD training fails to preserve adversarial robustness, and we propose KD with input gradient alignment (KDIGA) for remedy. Under certain assumptions, we prove that the student model using our proposed KDIGA can achieve at least the same certified robustness as the teacher model. Our experiments of KD contain a diverse set of teacher and student models with varying network architectures and sizes evaluated on ImageNet and CIFAR-10 datasets, including residual neural networks (ResNets) and vision transformers (ViTs). Our comprehensive analysis shows several novel insights that (1) With KDIGA, students can preserve or even exceed the adversarial robustness of the teacher model, even when their models have fundamentally different architectures; (2) KDIGA enables robustness transfer to pre-trained students, such as KD from an adversarially trained ResNet to a pre-trained ViT, without loss of clean accuracy; and (3) Our derived local linearity bounds for characterizing adversarial robustness in KD are consistent with the empirical results.

1 INTRODUCTION

Knowledge distillation (KD) (Hinton et al., 2015; Gou et al., 2021) is a popular machine learning framework for teacher-student training, with appealing applications to model compression in resource-constrained deep learning (Sun et al., 2019; Wang et al., 2019), such as memory-efficient inference on edge or mobile devices (Wang et al., 2021b; Lyu & Chen, 2020). In essence, KD trains a small model under the supervision of a large teacher model with the goal of improving or retaining the performance of the student model. For classification tasks, existing works mainly focus on preserving the accuracy of the teacher model (Zagoruyko & Komodakis, 2016b; Passban et al., 2020; Mirzadeh et al., 2020), while ignoring other important properties, such as adversarial robustness. For a student model, failing to preserve the same level of adversarial robustness as the teacher model can bring about a false sense of successful knowledge distillation when put into deployment. Therefore, ensuring and improving the adversarial robustness of the student model is critical to the safe deployment of the model in many practical applications.

To illustrate the critical but overlooked failure mode of standard KD, in Figure1 we show that it cannot preserve the adversarial robustness of the teacher model, and propose to use input gradient alignment in KD (we name it KDIGA) for better adversarial robustness preservation. In addition to empirical evidence, in this paper we also prove that our method can make the student achieve at least the same certified robustness as the teacher model under certain assumptions. When comparing our method with other baselines on ImageNet (Deng et al., 2009) and CIFAR-10 (Krizhevsky et al.) datasets, the results show substantial improvement in the adversarial robustness of the student models obtained by our method.

To demonstrate the generality of our proposed KDIGA method, we further study the transferability of adversarial robustness between convolutional neural networks (CNNs) (He et al., 2016) and vision

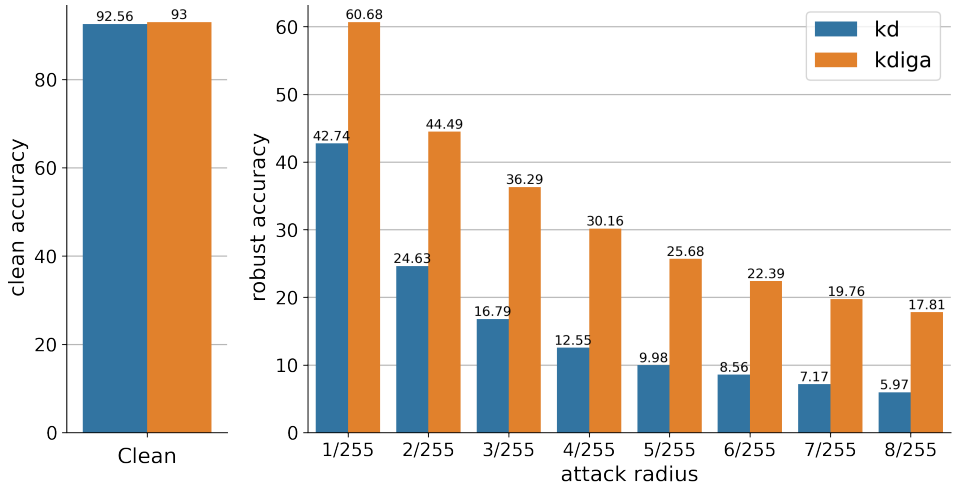

Figure 1: Clean accuracy (%) and robust accuracy (%) of the student models (MobileNetV2) against 20-step PGD attack (Madry et al., 2017) with different radii on CIFAR-10. The two students are distilled from the same adversarially trained WideResNet with TRADES (Zhang et al., 2019). "KD" stands for the standard knowledge distillation and "KDIGA" stands for our proposed knowledge distillation with input gradient alignment.

transformers (ViTs) (Dosovitskiy et al., 2020). We show that our method enables the transfer of adversarial robustness between these two fundamentally different architectures. We also show that KDIGA can improve the adversarial robustness of a pre-trained ViT without sacrificing the clean accuracy. We also extend our theoretical analysis and use local linearity measures to characterize the transfer of adversarial robustness in KD, and show that our derived performance bounds match the trends of the empirical robustness.

Our Contributions

- We propose to use KD with input gradient alignment (KDIGA) to train both accurate and adversarially robust student models in knowledge distillation. For instance, using KDIGA, the robust accuracy of the student model can be significantly increased from 5.97% to 17.81% compared with KD on CIFAR-10, with even better clean accuracy, as shown in Figure 1. On ImageNet, the robust accuracy of the student model is increased from 1.5% to 37.5% using KDIGA compared with KD.
- We show that adversarial robustness can be transferred between fundamentally different architectures with KDIGA, i.e., the robust accuracy of ResNet18 distilled from ViTs can achieve or even exceed the teacher's robust accuracy.
- KDIGA also extends to pre-trained student models. When we distill from adversarially trained ResNet50 to the normally pre-trained ViT in a fine-tuning approach, the robust accuracy of ViT boosts up to $11.1\times$ larger, together with even higher clean accuracy on the ImageNet dataset. We also find that students with higher learning capacity can achieve better results.
- We prove that the student model distilled with KDIGA can achieve at least the same certified robustness as the teacher with some mild assumptions. We further generalize the analysis and provide a bound with local linearity measures for characterizing adversarial robustness in KD, which is consistent with the empirical results on ImageNet and CIFAR-10.

2 Related Work

There are some recent works studying when and how adversarial robustness will transfer in different machine learning settings, such as transfer learning (Hendrycks et al., 2019; Chen et al., 2020; Shafahi et al., 2019), representation learning (Chan et al., 2020) and Model-agnostic meta-learning

(MAML) (Wang et al., 2021a). In contrast, we focus on the setting of knowledge distillation. The basic Knowledge Distillation (KD) formulates the supervised learning objective as

$$\arg\min_{f^s} \mathcal{L}_{KD}(\boldsymbol{x}, y) = \arg\min_{f^s} \lambda_{CE}\mathcal{L}_{CE}(f^s(\boldsymbol{x}), y) + \lambda_{KD}T^2\mathcal{L}_{KL}(f^s(\boldsymbol{x})/T, f^t(\boldsymbol{x})/T) \quad (1)$$

where f^s is the student model , f^t is the teacher model, $(\boldsymbol{x}, y) \in \mathcal{D}$, $\mathcal{D}$ is the training set, $\mathcal{L}_{CE}$ is the cross-entropy loss, $\mathcal{L}_{KL}$ is the KL-divergence loss, λ_{CE} and λ_{KD} are constant factors to balance the two losses, and T is a temperature factor.

One effective way to train adversarially robust model is adversarial training (Madry et al., 2017; Zhang et al., 2019; Engstrom et al., 2019), which adds adversarial perturbations to the inputs during training and forces the model to learn robust predictions. Goldblum et al. (2020) follows the same idea and formulates an adversarially robust distillation (ARD) objective using adversarial training:

$$\arg\min_{f^s} \mathcal{L}_{ARD}(\boldsymbol{x}, y) = \arg\min_{f^s} \lambda_{CE}\mathcal{L}_{CE}(f^s(\boldsymbol{x}), y) + \lambda_{KD}T^2\mathcal{L}_{KL}(f^s(\boldsymbol{x}+\delta)/T, f^t(\boldsymbol{x})/T), \quad (2)$$

$$\text{and} \quad \delta = \arg\max_{\|\delta\|_p \leq \epsilon} \mathcal{L}_{CE}(f^s(\boldsymbol{x} + \delta), y). \quad (3)$$

However, it is computationally expensive to calculate the adversarial perturbations for all training data especially when the dataset is large-scale (e.g. the ImageNet dataset (Deng et al., 2009)). There are some major differences between our method and ARD. Firstly, our method does not use adversarial training and thus is much more computationally efficient. Secondly, Our method can also be used together with ARD and further improve the robust accuracy of the student model.

Projected gradient descent (PGD) is one of the most commonly used adversarial attacks for both adversarial robustness evaluation and adversarial training, which solves Eq. 3 by iteratively taking gradient ascent by

$$\mathbf{x}_{t+1}^{adv} = Clip_{\mathbf{x}_0,\epsilon}(\mathbf{x}_t^{adv} + \alpha \cdot \text{sgn}\left(\nabla_{\mathbf{x}}\mathcal{L}_{CE}\left(\mathbf{x}_t^{adv}, y\right)\right)), \quad (4)$$

where $t = 1, \cdots, T$, T is the number of iterations, $\mathbf{x}_t^{adv}$ stands for the solution after t iterations, ∇_x denotes the gradient with respect to x, and $Clip_{\mathbf{x}_0,\epsilon}(\cdot)$ denotes clipping the values to make each $\mathbf{x}_{t+1}^{adv}$ within $[\mathbf{x}_0 - \epsilon, \mathbf{x}_0 + \epsilon]$, according to the ℓ_p norm bounded threat model. The adversarial perturbation is then obtained by $\delta_{\text{pgd}} = \mathbf{x}_T^{adv} - \mathbf{x}_0$. In addition, AutoAttack (Croce & Hein (2020)) is currently the strongest white-box attack which evaluates adversarial robustness with a parameter-free ensemble of diverse attacks.

3 Knowledge Distillation with Input Gradient Alignment

In this section, we first introduce our proposed framework of knowledge distillation with input gradient alignment (KDIGA) which we find is critical for adversarial robustness preservation in knowledge distillation. Then we prove that the student model can achieve at least the same certified robustness as the teacher model under two assumptions. We also give a general bound to analyze the factors that affect the transferability of adversarial robustness in KDIGA.

3.1 Problem Formulation

Suppose $f^s(\boldsymbol{x}) : \mathbb{R}^D \to \mathbb{R}^N$ is the student model and $f^t(\boldsymbol{x}) : \mathbb{R}^D \to \mathbb{R}^N$ is the teacher model, where D is the input dimension and N is the number of classes. In KDIGA, we force the student to learn both the logits and gradient knowledge from the teacher model, so the objective is defined as:

$$\begin{aligned}\arg\min_{f^s} \mathcal{L}_{IGA}(\boldsymbol{x}, y) = \arg\min_{f^s} \big[&\lambda_{CE}\mathcal{L}_{CE}(f^s(\boldsymbol{x}), y) + \lambda_{KL}T^2\mathcal{L}_{KL}(f^s(\boldsymbol{x})/T, f^t(\boldsymbol{x})/T) \\ &+\lambda_{IGA}\|\nabla_{\boldsymbol{x}}\mathcal{L}_{CE}(f^s(\boldsymbol{x}), y) - \nabla_{\boldsymbol{x}}\mathcal{L}_{CE}(f^t(\boldsymbol{x}), y\|_2)\big],\end{aligned} \quad (5)$$

where $(\boldsymbol{x}, y) \in \mathcal{D}$ is the input image and the corresponding label in the training dataset, $\mathcal{L}_{CE}$ and $\mathcal{L}_{KL}$ stand for the cross-entropy loss and the KL-divergence loss respectively, $\lambda_{CE}, \lambda_{KL}$ and λ_{IGA} are constants that balance the trade-off between different losses, T is the temperature factor, and $\|\cdot\|_2$ is the ℓ_2 norm. The pseudo code of KDIGA is shown in Algorithm 1.

Algorithm 1: Pseudocode of KDIGA

Input: teacher f^t, student f^s_θ with trainable parameters θ, training set $\mathcal{D}$,λ_{CE}, λ_{KL}, λ_{IGA}, learning rate η, number of epochs N_{epochs}.
Output: adversarially robust student f^s_θ.

```
for epoch ∈ N_epochs do
    for batch (x, y) ∈ D do
        p_s, p_t ← f^s_θ(x), f^t(x);
        ℓ_s, ℓ_t ← L_CE(p_s, y), L_CE(p_t, y);
        ℓ_KL ← T^2 L_KL(p_s/T, p_t/T);
        g_s, g_t ← ∇_x ℓ_s, ∇_x ℓ_t;
        ℓ_iga ← λ_CE ℓ_s + λ_KL ℓ_KL + λ_IGA ||g_s − g_t||_2;
        θ ← θ − η ∇_θ ℓ_iga;
    end
end
```

Besides, we show two ways to combine our method with adversarial training strategies for KD using ARD (Goldblum et al., 2020), i.e., KDIGA-ARD$_C$ and KDIGA-ARD$_A$. The objective for them are

$$\begin{aligned}\arg\min_{f^s} \mathcal{L}_{IGA_C}(\boldsymbol{x}, y) = \arg\min_{f^s} \big[&\lambda_{CE}\mathcal{L}_{CE}(f^s(\boldsymbol{x}), y) + \lambda_{KL}T^2\mathcal{L}_{KL}(f^s(\boldsymbol{x}+\delta)/T, f^t(\boldsymbol{x})/T) \\ &+\lambda_{IGA}\|\nabla_{\boldsymbol{x}}\mathcal{L}_{CE}(f^s(\boldsymbol{x}), y) - \nabla_{\boldsymbol{x}}\mathcal{L}_{CE}(f^t(\boldsymbol{x}), y\|_2)\big],\end{aligned} \tag{6}$$

$$\begin{aligned}\arg\min_{f^s} \mathcal{L}_{IGA_A}(\boldsymbol{x}, y) = \arg\min_{f^s} \big[&\lambda_{CE}\mathcal{L}_{CE}(f^s(\boldsymbol{x}), y) + \lambda_{KL}T^2\mathcal{L}_{KL}(f^s(\boldsymbol{x}+\delta)/T, f^t(\boldsymbol{x}+\delta)/T) \\ &+\lambda_{IGA}\|\nabla_{\boldsymbol{x}}\mathcal{L}_{CE}(f^s(\boldsymbol{x}+\delta), y) - \nabla_{\boldsymbol{x}}\mathcal{L}_{CE}(f^t(\boldsymbol{x}+\delta), y\|_2)\big],\end{aligned} \tag{7}$$

where "IGA_C" is in short for KDIGA-ARD$_C$ and "IGA_A" is in short for KDIGA-ARD$_A$, δ is calculated by Eq. 3 as inner maximization. KDIGA-ARD$_C$ is a direct combination of the original ARD formulation in Eq. 2 with our proposed IGA loss on clean samples as an additional regularization. KDIGA-ARD$_A$ further considers perturbed samples in IGA. Their key difference is that KDIGA-ARD$_C$ only aligns predictions of student on perturbed samples with the predictions of teacher on clean samples in the KL-divergence loss. On the other hand, KDIGA-ARD$_A$ forces the student to align with both the predictions and input gradients of the teacher on perturbed samples. We also tried other variants but did not observe notable differences.

3.2 Preservation of Certified Robustness

In this section, we prove that using KDIGA, the student model can provably achieve as good robustness as the the teacher model's in ideal situations. We first formally define δ-robust in Definition 1.

Definition 1. *(δ-robust) Classifier $f(\boldsymbol{x}) : \mathbb{R}^D \to \mathbb{R}^N$ is δ-robust if*

$$\arg\max f(\boldsymbol{x}+\boldsymbol{\epsilon}) = \arg\max f(\boldsymbol{x}), \quad \forall x \in \mathcal{D}, \forall \epsilon \in [\mathbf{0}, \boldsymbol{\delta}]^D. \tag{8}$$

Under mild assumptions, we aim to show that if the teacher model has a robust radius of δ, then the student model is at least δ-robust under ideal situations. The first assumption is the *perfect student* assumption in which we suppose $f^s : \mathbb{R}^D \to \mathbb{R}^N$ is a student model distilled from the teacher model $f^t : \mathbb{R}^D \to \mathbb{R}^N$ using distillation loss $\mathcal{L}$, and f^s is a perfect student if

$$\mathcal{L}(\boldsymbol{x}, y) = 0, \forall (\boldsymbol{x}, y) \in \mathcal{D}, \tag{9}$$

which means the student model trust the teacher and can perfectly learn the knowledge defined by the distillation objective. The second is the *local linearity* assumption, which assumes that neural networks with piece-wise linear activation functions are locally linear (Sattelberg et al. (2020); Croce et al. (2019); Lee et al. (2019)) and the certified robust area falls into these piece-wise linear regions. These two assumptions collaboratively build an ideal situation of knowledge distillation in which we can derive a strong property of KDIGA that the certified robustness of the student model can be as good as or even better than that of the teacher model. Proposition 1 concludes our statement.

Proposition 1. *Suppose the teacher model $f^t : \mathbb{R}^D \to \mathbb{R}^N$ is δ-robust, $f^s : \mathbb{R}^D \to \mathbb{R}^N$ is a perfect student trained using KDIGA, then f^s is at least δ-robust.*

We give a proof for Proposition 1 and illustrate why the knowledge distillation without input gradient alignment cannot preserve the adversarial robustness in Appendix A.

3.3 General Bound for the Adversarial Robustness of the Student Model

In this section, we derive a general bound for the adversarial robustness of the student model in knowledge distillation. No assumption is needed for this bound, and the knowledge distillation method is not limited to any specific one. To derive this bound, we first introduce the Local Linearity Measure (LLM, Qin et al. (2019)) in Definition 2.

Definition 2. *(Local Linearity Measure) The local linearity of a classifier $f(\boldsymbol{x}) : \mathbb{R}^D \to \mathbb{R}^N$ is measured by the maximum absolute difference between the cross-entropy loss and its first-order Taylor expansion in the δ-neighborhood:*

$$LLM(f, \boldsymbol{x}, \delta) = \max_{\boldsymbol{\epsilon} \in B(\delta)} \left| \mathcal{L}_{CE}(f(\boldsymbol{x} + \boldsymbol{\epsilon}) - \mathcal{L}_{CE}(f(\boldsymbol{x})) - \boldsymbol{\epsilon}^T \nabla_{\boldsymbol{x}} \mathcal{L}_{CE}(f(\boldsymbol{x})) \right|. \tag{10}$$

Proposition 2. *Consider a student model $f^s : \mathbb{R}^D \to \mathbb{R}^N$ distilled from a teacher model $f^t : \mathbb{R}^D \to \mathbb{R}^N$, then $\forall \boldsymbol{\epsilon} \in B(\delta)$,*

$$\left| \mathcal{L}_{CE}(f^s(\boldsymbol{x} + \boldsymbol{\epsilon}), y) - \mathcal{L}_{CE}(f^t(\boldsymbol{x} + \boldsymbol{\epsilon}), y) \right| \leq \gamma^s + \gamma^t + \phi \tag{11}$$

where $\gamma^s = LLM(f^s, \boldsymbol{x}, \delta)$, $\gamma^t = LLM(f^t, \boldsymbol{x}, \delta)$, and $\phi = \mathcal{L}_{CE}(f^s(\boldsymbol{x}), y) + \mathcal{L}_{CE}(f^t(\boldsymbol{x}), y) + \delta \| \nabla_{\boldsymbol{x}} \mathcal{L}_{CE}(f^s(\boldsymbol{x}), y) - \nabla_{\boldsymbol{x}} \mathcal{L}_{CE}(f^t(\boldsymbol{x}), y) \|$, and $\| \cdot \|$ is a norm.

The proof for Proposition 2 can be found in Appendix B.

Proposition 2 states that the adversarial robustness of the student model can be bounded by the LLM of both the student model and the teacher model, the clean accuracy of the student model, and the alignment of the student input gradient with the teacher input gradient. We will use this to further analyze the performance of different knowledge distillation methods in Section 4.

4 Experiments

In this section, the ImageNet (Deng et al., 2009) and CIFAR-10 (Krizhevsky et al.) datasets are used for model training and performance evaluation.

4.1 Settings

Teacher Models We use pre-trained and publicly available neural networks of varying architectures as teacher models. For the ImageNet dataset, we use both adversarially trained CNNs and normally trained vision transformers (ViTs) as the teacher models. We use the checkpoint of ResNet50 (He et al., 2016) provided by Engstrom et al. (2019) which is adversarially trained with an attack radius of $4/255$. We also incorporate ViTs (Dosovitskiy et al., 2020) as teacher models because they are shown to have better adversarial robustness than CNNs (Shao et al., 2021; Paul & Chen, 2021; Naseer et al., 2021), and we are interested in the transferability of adversarial robustness between different architectures. For the CIFAR-10 dataset, we use the WideResNet (Zagoruyko & Komodakis, 2016a) adversarially trained with TRADES (Zhang et al., 2019) as the teacher model.

Student Models For the ImageNet dataset, we mainly use ResNet18 (He et al., 2016) as the student model for experiments . To study the effect of model size, we also consider ResNet34, ResNet50 and ResNet101. In addition, we use ViT-S/16 (Dosovitskiy et al., 2020) as the student model to study the transferability of adversarial robustness from a CNN teacher to a ViT student. Unless specified, the student models are all trained from scratch. Because the training of ViT relies on the large-scale pre-training (Dosovitskiy et al., 2020), we use the pre-trained version provided by Wightman (2019) and apply the knowledge distillation methods as a fine-tuning process. For the CIFAR-10 dataset, we use MobileNetV2 (Sandler et al., 2018) as the student model.

Training Configurations For knowledge distillation on the ImageNet dataset, we run all distillation for 50 epochs with a batch size of 128, an initial learning rate of 0.1 for training from scratch and 0.00001 for fine-tuning, with milestones at $[20, 30, 40]$ of a decreasing rate of 0.1. The SGD optimizer with 0.9 momentum is used to update the model parameters, and a weight decay of 0.0001 is applied. For basic knowledge distillation, we set the temperature to 1 and the coefficients of the cross-entropy loss and KL-divergence loss both to 0.5. For KDIGA, we keep the same setting as that of the basic KD, and set the coefficient of the input gradient alignment term to $\frac{10^3}{B}$, where B is the batch size of the inputs.

For experiments on the CIFAR-10 dataset, we run distillation for 200 epochs with a batch size of 125, an initial learning rate of 0.1 with milestones at $[100, 150]$ of a decreasing rate of 0.1. The SGD optimizer with a momentum of 0.9 and a weight decay of 0.0002 is used to update the parameters. We set the coefficients of the cross-entropy loss and the KL-divergence loss both to 0.5, and the coefficient for the input gradient alignment to $\frac{10}{B}$, where B is the batch size.

Evaluation Metrics Using the test sets of ImageNet and CIFAR-10, we evaluate both the standard accuracy and the robust accuracy against adversarial attacks of the student models. We conduct ℓ_∞ norm bounded adversarial perturbations to generate adversarial examples for evaluating robust accuracy (the pixel value is scaled between 0 to 1), where we use a 40-step projected gradient descent (PGD) attack (Madry et al., 2017) and the parameter-free AutoAttack (Croce & Hein, 2020) for 1000 ImageNet test samples, and a 20-step PGD attack and the AutoAttack for all CIFAR-10 test samples. Different attack radii are used to test the robustness of the model under different degrees of adversarial perturbations.

Notation of Comparative Methods We denote the standard knowledge distillation method as "KD", knowledge distillation combined with adversarial training proposed by Goldblum et al. (2020) as "ARD", our method as "KDIGA", and the two kinds of combinations of KDIGA and ARD defined in Section 3.1 as "KDIGA-ARD$_C$" and "KDIGA-ARD$_A$". "Teacher $\xrightarrow{\text{Method}}$ Student" stands for the distillation from the "Teacher" to the "Student" using "Method".

4.2 Results on the ImageNet Dataset

We compare our method with the standard training (ST) and knowledge distillation (KD) on the ImageNet dataset. We find that adversarially robust distillation (ARD) proposed by Goldblum et al. (2020) cannot generalize to large-scale dataset, which shows no convergence in the setting as described in Section 4 with a very low training speed. So we only compare with ARD in the experiments on CIFAR-10 in Section 4.3. We run all experiments on ImageNet for 50 epochs to save training cost and expect better performance can be achieved with more training epochs, e.g. 100 epochs.

Table 1 shows the robust accuracy of the models trained using different training strategies, i.e., standard training (ST), standard knowledge distillation (KD) and knowledge distillation with input gradient alignment (KDIGA). Table 2 supplements some results of KDIGA against AutoAttack, as it is the strongest attack method in the current literature. We show the trend in the AutoAttack result is similar to that of PGD. From these results, we conclude the following observations.

Standard Knowledge Distillation Cannot Preserve Adversarial Robustness As shown in Table 1, models trained using standard training are vulnerable to adversarial perturbations. The standard knowledge distillation shows no preservation of adversarial robustness from teacher models, e.g., ResNet18 distilled from ViT-S/16 and the adversarially trained ResNet50 still have low robust accuracy against PGD attack with various radii. When the PGD attack radius is 0.003, the robust accuracy of the student distilled from {ResNet50 (AT), ViT-S/16} is only $\{1.5\%, 2.9\%\}$ compared with the teacher's $\{59.4\%, 24.6\%\}$, where "AT" is in short for adversarial training. For the attack radius of 0.005, the robust accuracy of the student distilled from {ResNet50 (AT), ViT-S/16} is only $\{0.0\%, 0.6\%\}$ compared with the teacher's $\{57.2\%, 10.2\%\}$.

Input Gradient Alignment Makes Students More Robust From Table 1 and Table 2, students distilled using KDIGA have higher robust accuracy than those distilled with KD or trained with ST. When the teacher model is ResNet50(AT) and the PGD attack radius is 0.003, ResNet18 distilled with KDIGA has a robust accuracy of 37.5%, while the ResNet18 distilled with KD only has a

Table 1: Robust accuracy (%) of student models against 40-step PGD attack with different radii and clean accuracy (%) on the ImageNet dataset. Robust accuracy of the teacher models are shown in brackets. The pre-trained student model is denoted with "*" where the distillation is conducted as a fine-tuning process. Other students are all trained from scratch. "ST" means the model is trained following the standard approach without distillation nor adversarial training. "AT" means the model is obtained by adversarial training.

		PGD Attack radius			
Model	**Clean**	**0.001**	**0.003**	**0.005**	**0.01**
ResNet18 (ST)	68.7 (-)	24.9 (-)	2.0 (-)	0.6 (-)	0.0 (-)
ViT-S/16 (ST)	77.6 (-)	55.4 (-)	24.6 (-)	10.2 (-)	1.0 (-)
ViT-S/16 (ST) $\xrightarrow{\text{KD}}$ ResNet18	69.0 (77.6)	30.1 (55.4)	2.9 (24.6)	0.6 (10.2)	0.0 (1.0)
ViT-S/16 (ST) $\xrightarrow{\text{KDIGA}}$ ResNet18	60.0 (77.6)	51.0 (55.4)	32.7 (24.6)	18.0 (10.2)	3.3 (1.0)
ViT-B/16 (ST) $\xrightarrow{\text{KDIGA}}$ ResNet18	64.7 (76.3)	52.8 (48.9)	26.6 (14.6)	11.3 (6.0)	0.7 (0.9)
ViT-L/16 (ST) $\xrightarrow{\text{KDIGA}}$ ResNet18	65.9 (80.0)	53.2 (55.1)	28.6 (23.4)	12.4 (9.9)	1.4 (1.8)
DEIT-S/16 (ST) $\xrightarrow{\text{KDIGA}}$ ResNet18	63.6 (77.7)	53.1 (48.9)	31.5 (17.6)	15.6 (7.1)	1.6 (1.1)
ResNet50 (AT) $\xrightarrow{\text{KD}}$ ResNet18	66.3 (63.1)	25.7 (61.9)	1.5 (59.4)	0.0 (57.2)	0.0 (49.0)
ResNet50 (AT) $\xrightarrow{\text{KDIGA}}$ ResNet18	54.2 (63.1)	48.2 (61.9)	37.5 (59.4)	26.5 (57.2)	9.2 (49.0)
ResNet50 (AT) $\xrightarrow{\text{KDIGA}}$ ResNet34	59.2 (63.1)	53.9 (61.9)	42.7 (59.4)	31.2 (57.2)	12.1 (49.0)
ResNet50 (AT) $\xrightarrow{\text{KDIGA}}$ ResNet50	58.8 (63.1)	53.7 (61.9)	42.2 (59.4)	31.9 (57.2)	12.4 (49.0)
ResNet50 (AT) $\xrightarrow{\text{KDIGA}}$ ResNet101	60.3 (63.1)	55.3 (61.9)	44.7 (59.4)	33.1 (57.2)	12.7 (49.0)
ResNet50 (AT) $\xrightarrow{\text{KDIGA}}$ ViT-S/16*	77.7 (63.1)	65.3 (61.9)	50.4 (59.4)	33.5 (57.2)	11.1 (49.0)

Table 2: Robust accuracy (%) of student models against AutoAttack with different radii and clean accuracy (%) on the ImageNet dataset. Robust accuracy of the teacher models are shown in brackets. The pre-trained student model is denoted with "*" where the distillation is conducted as a fine-tuning process. Other students are all trained from scratch. "ST" means the model is trained following the standard approach without distillation nor adversarial training. "AT" means the model is obtained by adversarial training.

		AutoAttack Attack radius			
Model	**Clean**	**0.001**	**0.003**	**0.005**	**0.01**
ResNet18 (ST)	68.7 (-)	14.3 (-)	0.4 (-)	0.0 (-)	0.0 (-)
ViT-S/16 (ST)	77.6 (-)	48.1 (-)	6.0 (-)	0.5 (-)	0.0 (-)
ViT-S/16 (ST) $\xrightarrow{\text{KDIGA}}$ ResNet18	60.0 (77.6)	47.2 (48.1)	25.0 (6.0)	10.1 (0.5)	0.7 (0.0)
ViT-B/16 (ST) $\xrightarrow{\text{KDIGA}}$ ResNet18	64.7 (76.3)	49.6 (39.8)	19.4 (5.4)	5.0 (0.6)	0.0 (0.0)
ViT-L/16 (ST) $\xrightarrow{\text{KDIGA}}$ ResNet18	65.9 (80.1)	49.6 (46.6)	19.1 (8.5)	5.8 (1.0)	0.0 (0.0)
DEIT-S/16 (ST) $\xrightarrow{\text{KDIGA}}$ ResNet18	63.6 (80.1)	50.0 (0.4)	23.7 (0.0)	7.8 (0.0)	0.1 (0.0)
ResNet50 (AT) $\xrightarrow{\text{KDIGA}}$ ResNet18	54.2 (63.1)	45.9 (47.5)	31.9 (42.5)	19.1 (35.0)	3.9 (30.0)
ResNet50 (AT) $\xrightarrow{\text{KDIGA}}$ ViT-S/16*	77.7 (63.1)	65.3 (47.5)	32.6 (42.5)	13.4 (35.0)	1.1 (30.0)

robust accuracy of 1.5%, and ResNet18 (ST) only has a robust accuracy of 2.0%. When the teacher model is ViT-S/16 (ST) and the PGD attack radius is 0.003, ResNet18 distilled with KDIGA has a robust accuracy of 32.7%, while the ResNet18 distilled with KD only has a robust accuracy of 2.9%. This shows that our proposed input gradient alignment plays the key role to help preserve the adversarial robustness during knowledge distillation.

Adversarial Robustness Can Transfer Between CNNs and Vision Transformers Vision transformers and CNNs have entirely different model architectures, while Table 1 and Table 2 show that the adversarial robustness can be transferred between them with KDIGA. We have already shown that the adversarial robustness can be transferred from ViTs to CNNs in the previous paragraph, and here we show the reverse also holds. With PGD attack radius of $\{0.003, 0.005, 0.1\}$, the robust accuracy of ViT/S-16 obtained by ST is only $\{24.6\%, 10.2\%, 1.0\%\}$ while the distilled ViT-S/16 has an accuracy of $\{50.4\%, 33.5\%, 11.1\%\}$. While under AutoAttack, when the attack radius is

$\{0.001, 0.003, 0.005\}$, the robust accuracy of ViT-S/16 (ST) is $\{14.3\%, 0.4\%, 0.0\%\}$, but after distillation, the robust accuracy becomes $\{65.3\%, 32.6\%, 13.4\%\}$. As ViTs are difficult to train even in the standard setting, this result shows that we can consider to transfer adversarial robustness from adversarially trained CNNs to ViTs to obtain robust ViTs.

Input gradient alignment Works for Pre-trained Models The student model in standard knowledge distillation is generally trained from scratch. While in the experiments of "ResNet50 $\xrightarrow{KDIGA}$ ViT-S/16*" as shown in Table 1 and Table 2, we take the pre-trained ViT as the student to help the training converge in a shorter time. In Table 2, ViT-S/16 remains the high clean accuracy of 77.7% after distillation compared with the original clean accuracy of 77.6% in standard pre-training. This result shows the feasibility to further promote the adversarial robustness of a pre-trained model using our proposed input gradient alignment without harming the model's performance on the clean dataset in knowledge distillation, which gives a novel and inspiring approach to train new robust models more efficiently at less cost of the clean accuracy.

Students Can Obtain Even Better Adversarial Robustness Than Teachers From Proposition 1, we proved that the student model can achieve at least the same certified robustness as the teacher model's under certain assumptions. The results in Table 1 and Table 2 also show that the student can obtain even higher robust accuracy against adversarial perturbations than teacher's with KDIGA empirically. For example, when the teacher model is not robust, i.e. DEIT-S/16, ResNet18 distilled from it with KDIGA achieves an robust accuracy of $\{50.0\%, 23.7\%, 7.8\%\}$ against AutoAttack with attack radius of $\{0.001, 0.003, 0.005\}$, while the teacher model only has an accuracy of $\{0.4\%, 0.0\%.0.0\%\}$ in the same situations. This shows that input gradient alignment can still help the student obtain better adversarial robustness even when the teacher is not very robust.

Students with Higher Learning Capacity Achieve Better Results In Table 1, we set the student model to different sizes, i.e., ResNet18, ResNet34, ResNet50, and ResNet101, to check the effect of the student's learning capacity on knowledge distillation. As shown in the table, both the clean accuracy and the robust accuracy increase as the the model size grows, e.g., ResNet101 achieves a robust accuracy of 12.7% under attack radius of 0.01 while ResNet18 achieves 9.2% in the same case. Therefore, we can expect students with higher learning capacities to achieve better performance. It is also worth noting that the student model in KD is commonly smaller than the teacher for some practical purposes like model compression. While in our setting, student can be larger than the teacher model (like the case of noisy student training), meaning we can train a robust teacher with less computing cost and then transfer the adversarial robustness to larger students, in order to eventually obtain a high-capacity and robust model.

4.3 Results on the CIFAR-10 Dataset

We compare our method with ST, KD and ARD on the CIFAR-10 dataset in Table 3. We show that KDIGA has the highest clean accuracy compared with other baseline methods. We also show that two combinations of our method with ARD have the highest robust accuracy in knowledge distillation.

We use the fair comparison setting described in Section 4.1, though we find ARD can achieve higher robust accuracy with λ_{CE} set to 0 and λ_{KL} set to 1, in which case its clean accuracy will decrease a lot. λ_{CE} is critical to preserve the clean accuracy and it is an important term in the standard KD. So we set both λ_{CE} and λ_{KL} to 0.5 as in the standard KD setting for all experiments.

From Table 3, we find the student distilled with KDIGA achieves comparable robust accuracy with ARD. But KDIGA has the advantage of dispensing the computation of the adversarial examples and thus is more cheap and efficient. Moreover, KDIGA has the highest clean accuracy which even exceeds the result of KD. When combined with KDIGA, both the KDIGA-ARD$_{\text{C}}$ and KDIGA-ARD$_{\text{A}}$ obtain higher robust accuracy than ARD. This shows that it is feasible to combine KDIGA with other state-of-the-art methods to further improve the student's robust accuracy.

Table 3: Robust accuracy (%) of student models against 20-step PGD attack with different radii and clean accuracy (%) on the CIFAR-10 dataset. All students are trained from scratch. "ST" means the model is trained following the standard approach without distillation nor adversarial training. "TRADES" means the model is adversarially trained using TRADES (Zhang et al., 2019).

		PGD Attack radius							
Model	**Clean**	**1/255**	**2/255**	**3/255**	**4/255**	**5/255**	**6/255**	**7/255**	**8/255**
WideResNet(TRADES)	84.92	82.36	79.35	75.99	72.28	68.54	64.54	60.57	56.68
MobileNetV2 (ST)	91.82	7.43	1.11	0.04	0.0	0.0	0.0	0.0	0.0
WideResNet(TRADES) $\xrightarrow{\text{KD}}$ MobileNetV2	92.56	42.74	24.63	16.79	12.55	9.98	8.56	7.17	5.97
WideResNet(TRADES) $\xrightarrow{\text{ARD}}$ MobileNetV2	91.65	80.35	68.34	58.43	49.49	41.74	33.78	26.17	20.73
WideResNet(TRADES) $\xrightarrow{\text{KDIGA}}$ MobileNetV2	**93.03**	60.68	44.49	36.29	30.16	25.68	22.39	19.76	17.81
WideResNet(TRADES) $\xrightarrow{\text{KDIGA-ARD}_\text{C}}$ MobileNetV2	92.22	83.29	71.76	61.85	53.46	45.28	37.70	31.12	**25.85**
WideResNet(TRADES) $\xrightarrow{\text{KDIGA-ARD}_\text{A}}$ MobileNetV2	90.67	81.82	70.57	60.69	52.58	45.50	38.59	32.72	**27.50**

Table 4: Bounds for adversarial robustness (as defined in Proposition 2) of different models on CIFAR-10. llm_ϵ is defined by Definition 2 where ϵ is the radius of perturbations. l_{CE} is the cross-entropy loss. $\|g^s - g^t\|_2$ calculates the l_2-norm of input gradient alignment term. "ST" means the model is trained following the standard approach without distillation nor adversarial training. "TRADES" means the model is adversarially trained using TRADES (Zhang et al., 2019). We calculate $\|g^s - g^t\|_2$ with WideResNet(TRADES) as the teacher model for MobileNetV2 (ST) for comparison, though the training process of MobileNetV2 (ST) doesn't involve a teacher model.

Model	$llm_{4/255}$	$llm_{8/255}$	l_{CE}	$\|g^s - g^t\|_2$
MobileNetV2 (ST)	12.413	21.691	0.364	4.099
WideResNet(TRADES) $\xrightarrow{\text{KD}}$ MobileNetV2	5.960	10.286	**0.218**	1.958
WideResNet(TRADES) $\xrightarrow{\text{ARD}}$ MobileNetV2	1.326	3.034	0.261	0.569
WideResNet(TRADES) $\xrightarrow{\text{KDIGA}}$ MobileNetV2	2.561	4.914	0.235	0.587
WideResNet(TRADES) $\xrightarrow{\text{KDIGA-ARD}_\text{C}}$ MobileNetV2	**1.081**	**2.421**	0.228	**0.339**
WideResNet(TRADES) $\xrightarrow{\text{KDIGA-ARD}_\text{A}}$ MobileNetV2	1.107	2.442	0.285	0.377

4.4 Local Linearity Bounds for Adversarial Robustness in Knowledge Distillation

From Proposition 2, we prove that the certified robustness of the student model can be bounded by the LLM (as defined in Definition 2), the cross-entropy loss, and the gradient alignment norm, if we regard other terms of the teacher as fixed. Table 4 shows the bounds for adversarial robustness of models trained on CIFAR-10. We randomly sample 1000 test samples to calculate the terms in the bounds.

In reference to Table 3 and Table 4, the empirical performance matches the theoretical insights that models with better adversarial robustness have smaller values in the bounds, i.e., MobileNetV2 trained with standard training has the highest bounds, and students distilled with KDIGA-ARD has the lowest bounds. The LLM bound and input gradient alignment norm for ARD are much lower than KD, showing that adversarial training also has the effect of improving the local linearity and aligning the input gradients. KDIGA achieves similar bounds as ARD though the training process does not use adversarial examples. Table 4 also shows that combining our method with ARD can further reduce the bounds and induce better adversarial robustness. KD only has the lowest cross-entropy loss while other terms are high, which can explain its failure in preserving adversarial robustness, as its objective design only focuses on improving standard accuracy.

5 Conclusion

This paper provides a comprehensive study on how and when can adversarial robustness transfer from the teacher model to student model in knowledge distillation, in addition to standard accuracy. For the *how*, we show that standard knowledge distillation fails to preserve adversarial robustness, and we propose a novel input gradient alignment technique (KDIGA) to address this issue. For the

when, under certain assumptions we prove that using KDIGA the student model can be at least as robust as the teacher model, and we generalize our theoretical analysis using local linearity measures. The superior performance of KDIGA over baselines in terms of improved adversarial robustness while retaining clean accuracy is empirically validated using CNNs and vision transformers.

REFERENCES

Alvin Chan, Yi Tay, and Yew-Soon Ong. What it thinks is important is important: Robustness transfers through input gradients. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 332–341, 2020.

Tianlong Chen, Sijia Liu, Shiyu Chang, Yu Cheng, Lisa Amini, and Zhangyang Wang. Adversarial robustness: From self-supervised pre-training to fine-tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 699–708, 2020.

Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *International Conference on Machine Learning*, pp. 2206–2216. PMLR, 2020.

Francesco Croce, Maksym Andriushchenko, and Matthias Hein. Provable robustness of relu networks via maximization of linear regions. In *the 22nd International Conference on Artificial Intelligence and Statistics*, pp. 2057–2066. PMLR, 2019.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Logan Engstrom, Andrew Ilyas, Hadi Salman, Shibani Santurkar, and Dimitris Tsipras. Robustness (python library), 2019. URL https://github.com/MadryLab/robustness.

Micah Goldblum, Liam Fowl, Soheil Feizi, and Tom Goldstein. Adversarially robust distillation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 3996–4003, 2020.

Jianping Gou, Baosheng Yu, Stephen J Maybank, and Dacheng Tao. Knowledge distillation: A survey. *International Journal of Computer Vision*, 129(6):1789–1819, 2021.

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*, pp. 770–778, 2016.

Dan Hendrycks, Kimin Lee, and Mantas Mazeika. Using pre-training can improve model robustness and uncertainty. In *International Conference on Machine Learning*, pp. 2712–2721. PMLR, 2019.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton. Cifar-10 (canadian institute for advanced research). URL http://www.cs.toronto.edu/~kriz/cifar.html.

Guang-He Lee, David Alvarez-Melis, and Tommi S Jaakkola. Towards robust, locally linear deep networks. *arXiv preprint arXiv:1907.03207*, 2019.

Lingjuan Lyu and Chi-Hua Chen. Differentially private knowledge distillation for mobile analytics. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 1809–1812, 2020.

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.

Seyed Iman Mirzadeh, Mehrdad Farajtabar, Ang Li, Nir Levine, Akihiro Matsukawa, and Hassan Ghasemzadeh. Improved knowledge distillation via teacher assistant. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 5191–5198, 2020.

Muzammal Naseer, Kanchana Ranasinghe, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Intriguing properties of vision transformers. *arXiv preprint arXiv:2105.10497*, 2021.

Peyman Passban, Yimeng Wu, Mehdi Rezagholizadeh, and Qun Liu. Alp-kd: Attention-based layer projection for knowledge distillation. *arXiv preprint arXiv:2012.14022*, 2020.

Sayak Paul and Pin-Yu Chen. Vision transformers are robust learners. *arXiv preprint arXiv:2105.07581*, 2021.

Chongli Qin, James Martens, Sven Gowal, Dilip Krishnan, Krishnamurthy Dvijotham, Alhussein Fawzi, Soham De, Robert Stanforth, and Pushmeet Kohli. Adversarial robustness through local linearization. *arXiv preprint arXiv:1907.02610*, 2019.

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*, pp. 4510–4520, 2018.

Ben Sattelberg, Renzo Cavalieri, Michael Kirby, Chris Peterson, and Ross Beveridge. Locally linear attributes of relu neural networks. *arXiv preprint arXiv:2012.01940*, 2020.

Ali Shafahi, Parsa Saadatpanah, Chen Zhu, Amin Ghiasi, Christoph Studer, David Jacobs, and Tom Goldstein. Adversarially robust transfer learning. *arXiv preprint arXiv:1905.08232*, 2019.

Rulin Shao, Zhouxing Shi, Jinfeng Yi, Pin-Yu Chen, and Cho-Jui Hsieh. On the adversarial robustness of visual transformers. *arXiv preprint arXiv:2103.15670*, 2021.

Siqi Sun, Yu Cheng, Zhe Gan, and Jingjing Liu. Patient knowledge distillation for bert model compression. *arXiv preprint arXiv:1908.09355*, 2019.

Ji Wang, Weidong Bao, Lichao Sun, Xiaomin Zhu, Bokai Cao, and S Yu Philip. Private model compression via knowledge distillation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 1190–1197, 2019.

Ren Wang, Kaidi Xu, Sijia Liu, Pin-Yu Chen, Tsui-Wei Weng, Chuang Gan, and Meng Wang. On fast adversarial robustness adaptation in model-agnostic meta-learning. In *International Conference on Learning Representations*, 2021a. URL https://openreview.net/forum?id=o81ZyBCojoA.

Yiran Wang, Xingyi Li, Min Shi, Ke Xian, and Zhiguo Cao. Knowledge distillation for fast and accurate monocular depth estimation on mobile devices. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2457–2465, 2021b.

Ross Wightman. Pytorch image models. https://github.com/rwightman/pytorch-image-models, 2019.

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

Sergey Zagoruyko and Nikos Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. *arXiv preprint arXiv:1612.03928*, 2016b.

Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael Jordan. Theoretically principled trade-off between robustness and accuracy. In *International Conference on Machine Learning*, pp. 7472–7482. PMLR, 2019.