Revisiting Adversarial Robustness Distillation: Robust Soft Labels Make Student Better

摘要

对抗训练是一种有效训练防御对抗攻击的模型的方法。尽管能够带来可靠的鲁棒性，但对抗训练普遍需要高容量的模型，即模型越大鲁棒性越好。这限制了对抗训练在小模型上的效果，尤其在某些存储和计算资源有限的场景下。本论文中，作者提出一种以预训练的大模型通过蒸馏来提升小模型鲁棒性的方法。作者从知识蒸馏的视角重新审视几个最先进的对抗训练方法，并确定提升鲁棒性的共同点：使用鲁棒性软标签。在这个观察下，作者提出了一种新的对抗鲁棒蒸馏方法，称为鲁棒性软标签对抗蒸馏（RSLAD）。RSLAD 充分利用了鲁棒性软标签来监督学生模型在自然和对抗样本的损失函数训练。作者经验性地证明了 RSLAD 在防御现有最先进的对抗攻击的有效性，以及在对抗鲁棒性蒸馏上提出一套理解 RSLAD 和鲁棒性软标签的方法。

关键词：对抗训练；鲁棒性软标签；对抗鲁棒蒸馏

1 引言

深度神经网络（DNNs）已经成为了解决真实世界复杂学习问题的标准模型，例如图像分类、语音识别和自然语言处理。然而，研究表明 DNNs 在对抗样本上表现出脆弱性，这些对抗样本是通过在输入样本上添加难以察觉的细微扰动而生成。这引起了关于 DNNs 应用方面的安全担忧，尤其是在自动驾驶和医学诊断等安全尤为关键的场景中。

目前已经证实了 DNNs 有不同类型的方法能够有效防御对抗攻击，其中对抗训练是最为有效的方法。对抗训练被视作是一种数据增强的技巧，它通过针对自然样本来生成不同的对抗样本训练模型。对抗训练一般被表述为最小最大优化问题，即通过最大化内部函数来生成对抗样本，同时最小化外部函数优化模型在该对抗样本上的参数。

虽然对抗训练带来了可靠的鲁棒性，但是仍存在需要限制在一些特定的应用场景中使用。可论证的是，最显著的缺点则是需要大容量的模型，即模型越大鲁棒性越好。然而，在许多场景中更偏好于轻小的模型，而非大模型。例如，小型 DNNs 应用在智能手机和自动驾驶车辆中，才能够承载有限容量和计算能耗。这激发了知识蒸馏在对抗训练上的应用，通过知识蒸馏提升小模型的鲁棒性，即对抗鲁棒蒸馏方法。

在本论文中，作者建立在过往的对抗训练和对抗鲁棒蒸馏工作上，通过对比几个最先进对抗训练模型的损失函数，发现通过蒸馏提升小型 DNNs 鲁棒性的关键因素：使用对抗训练模型的预测标签。这是一种使用鲁棒性软标签作为监督学习的方式。通过对比原始硬标签，鲁

棒性软标签更能呈现出教师模型的鲁棒性行为，提供更多的鲁棒信息来指导学生。这个观察激发了作者，设计出一种新的对抗鲁棒蒸馏方法来充分利用鲁棒性软标签。

2 相关工作

2.1 对抗攻击

给定 DNN 模型已知的参数，对抗样本能够通过 FGSM、PGD、CW 等方法生成。这些对抗攻击被应用于生成更为可靠的防御模型的对抗鲁棒评估方法。这些方法有效避免了不合理的防御模型的巧妙梯度遮挡或混淆作用。AA 攻击是由 PGD [3]、DLR、FAB-Attack 和黑盒 Square Attack 四种攻击方式组装而成的。目前 AA 攻击是理论上最强的对抗攻击方式。

2.2 对抗训练

对抗训练被认为是最有效的防御对抗样本的训练方式。最近，有大量关于这个领域的理解和方法生成。对抗训练被表述为如下的最小最大优化问题：

$$\begin{aligned} &\underbrace{arg\ \min_{\theta} \mathcal{L}_{min}\left(f\left(x', \theta\right), y\right)}_{Outer\ \ minimization} \\ &where\ x' = \underbrace{arg\ \max_{\|x'-x\|_p \leq \epsilon} \mathcal{L}_{max}\left(f\left(x', \theta\right), y\right)}_{Inner\ \ maximization} \end{aligned} \tag{1}$$

在标准对抗训练中，损失项 L_{min} 和 L_{max} 设置为相同损失函数，即最为常用的交叉熵损失函数。在内部最大化问题由 PGD 攻击解决。

大量提升标准对抗训练的有效性方法已经提出，包括使用更大更宽的模型、添加未标注数据、通过 KL 散度对鲁棒性和准确度进行权衡的理论（TRADES）、强化错误分类的训练方法（MART）、对抗权重干扰等。总而言之，在这些工作中，对鲁棒性有贡献的有更大模型、更多数据和使用 KL 损失作为内部最大化函数优化。

对抗训练方法并非完美，在现有的方法中最显著的缺点就是模型越小鲁棒性越差。通常来说，尽管大模型如 WideResNet-34-10 和 WideResNet-70-16 能够带来可观的鲁棒性提升，但是小模型如 ResNet-18 和 MobileNetV2 却是非常难以提升鲁棒性。这意味着，在手机设备、自动驾驶车辆和无人机等有限存储和计算资源的场景下，会限制它们的有效性。在本论文中，作者利用知识蒸馏技术提升小模型的鲁棒性，以及提升了现有对抗鲁棒性蒸馏的方法。

2.3 知识蒸馏

知识蒸馏是一种深度神经网络模型压缩方法 [2]，将大型的教师模型的知识萃取到轻小型的学生模型中。给定预训练好的教师模型 T，通过以下优化问题来蒸馏训练学生模型 S：

$$arg\ \min_{\theta}(1-\alpha)\mathcal{L}(S(x), y) + \alpha\tau^2 KL(S^{\tau}(x), T^{\tau}(x)), \tag{2}$$

知识蒸馏方法应用广泛，在噪声标注学习、AI 安全和自然语言处理等各种学习任务中均有体现。显而易见的是，最近几年自蒸馏这个分支吸引了很多的注意力。与传统的蒸馏方式不同，

Method	$\mathcal{L}_{\min}$	$\mathcal{L}_{\max}$	Student/Teacher
SAT	$\mathrm{CE}(f(x'), y)$	$\mathrm{CE}(f(x'), y)$	-
TRADES	$\mathrm{CE}(f(x), y) + \lambda \mathrm{KL}(f(x'), f(x))$	$\mathrm{KL}(f(x'), f(x))$	S: $f(\cdot)$; T: $f(\cdot)$
MART	$\mathrm{BCE}(f(x'), y) + \lambda(1 - f_y(x))\mathrm{KL}(f(x'), f(x))$	$\mathrm{CE}(f(x'), y)$	S: $f(\cdot)$; T: $f(\cdot)$
ARD	$(1-\alpha)\mathrm{CE}(S^\tau(x), y) + \alpha\tau^2 \mathrm{KL}(S^\tau(x'), T^\tau(x))$	$\mathrm{CE}(S(x'), y)$	S: $S(\cdot)$; T: $T(\cdot)$
IAD	$T_y(x')^\beta \mathrm{KL}(S^\tau(x'), T^\tau(x)) + (1 - T_y(x')^\beta)\mathrm{KL}(S^\tau(x'), S^\tau(x))$	$\mathrm{CE}(S(x'), y)$	S: $S(\cdot)$; T: $T(\cdot)$
RSLAD (ours)	$(1-\alpha)\mathrm{KL}(S(x), T(x)) + \alpha \mathrm{KL}(S(x'), T(x))$	$\mathrm{KL}(S(x'), T(x))$	S: $S(\cdot)$; T: $T(\cdot)$

表 1. 对抗训练方法损失函数总结

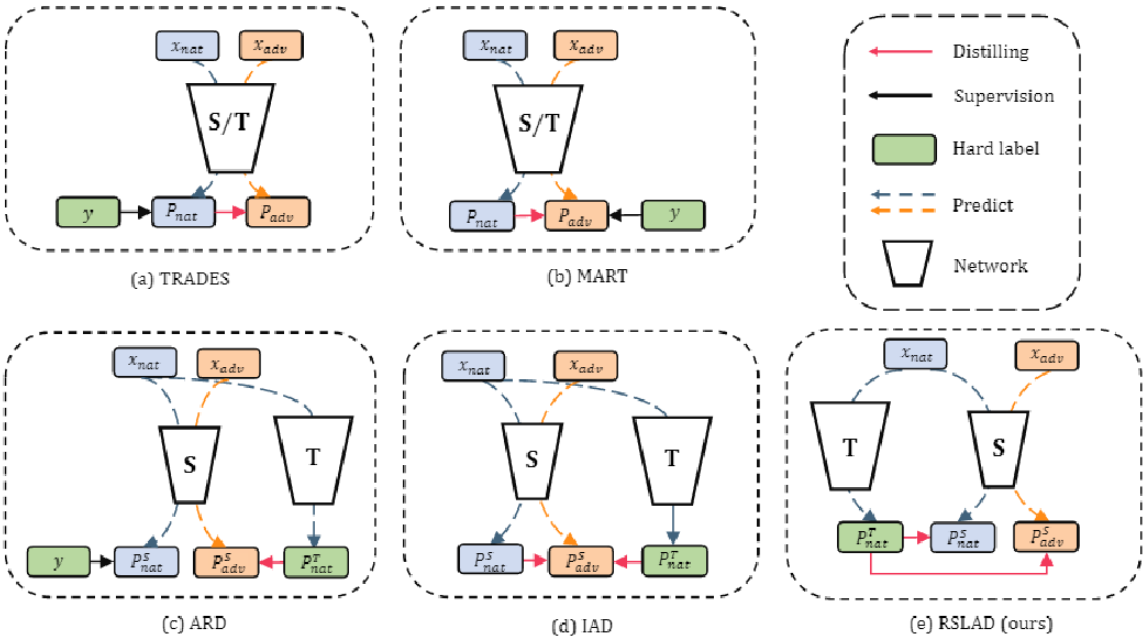

图 1. 对抗训练方法框架对比

自蒸馏是自身教导自己，而非是分开的教师模型。知识蒸馏同样已经应用到了对抗训练，通过预训练具备鲁棒性的教师模型以提升学生模型的鲁棒性。其中，教师模型是大型且更具鲁棒性的模型，或者是与学生模型有相同结构的模型。它会比从零开始训练一个学生模型的方式更具有鲁棒性，比如 ARD [1] 和 IAD 方式，说明鲁棒性特征也能够通过知识蒸馏的方式传递给学生模型。本论文中，作者在这些工作的基础上，提出了一种更有效果的对抗鲁棒蒸馏方法，以提升小型学生模型鲁棒性。

3 本文方法

3.1 本文方法概述

作者从知识蒸馏的角度重新审视最先进的对抗训练和对抗鲁棒蒸馏方法，通过比较几种最先进的 AT 方法所采用的损失函数 1，确定并提出了一个能够提升鲁棒性的共同因素：使用对抗训练模型的预测。作者对四种对抗训练方法和两种对抗鲁棒蒸馏方法研究 1，提出使用鲁棒性软标签的新对抗鲁棒蒸馏方法 RSLAD [4]。

3.2 RSLAD 方法

跟其他对抗训练方法不同，RSLAD 方法的学生模型使用由教师模型产生的鲁棒性软标签，以此来监督训练所有自然和对抗样本的损失项。相比硬标签，鲁棒性软标签能够提供更多的鲁棒信息来指导学生模型的学习。在 RSLAD 方法中，通过 KL 散度损失函数来充分利用鲁棒性软标签的信息来提升学生模型的鲁棒性。使用教师模型产生的鲁棒性软标签来监督学生模型在所有损失条件下对自然和对抗样本的训练，即 RSLAD 中没有使用原始的硬标签 y。RSLAD 的优化函数如下：

$$\begin{aligned} &\underset{\theta_S}{arg\ min}(1-\alpha)KL(S(x),T(x))+\alpha KL(S(x'),T(x)) \\ &where\ x'=\underset{\|x'-x\|_p\leq\epsilon}{arg\ max}KL(S(x),T(x)) \end{aligned} \tag{3}$$

鲁棒性软标签由对抗训练过的教师模型 $T(x)$ 生成，用于监督学生模型干净样本的外部最小化函数。其中关键点是将常用的 CE 损失换成了 KL 散度损失，以此衡量两个模型输出分布之间的差异。

4 复现细节

4.1 与已有开源代码对比

作者在本文中提出的对抗鲁棒蒸馏方法 RSLAD，使用预训练的教师模型 WideResNet-34-10 和 WideResNet-70-16，通过蒸馏方式分别训练两个小模型 ResNet-18 和 MobileNetV2，成功提升了两个小模型的鲁棒性，结果都优于最先进的对抗训练方法。与此同时，针对不同的标签在模型上的作用进行了分析，得到了鲁棒性软标签更能提升模型鲁棒性的结论。在本次论文复现中，通过学习本篇论文的思路和参考作者公开的源代码，实现了作者所提供的 RSLAD 方法。

4.2 实验环境搭建

实验环境：python3.9

数据集：cifar10

5 实验结果分析

在训练学生模型 ResNet-18 和 MobileNetV2 中，使用数据集 cifar10，以及通过 PGD 对抗攻击算法生成对抗样本，并且结合自然样本一同训练。

使用不同的对抗攻击算法，分别对两个训练好的学生模型 ResNet-18 和 MobileNetV2 进行鲁棒测试 2 3。通过对比表 2 和表 3 的结果，可知本次复现的结果基本与作者的实验结果保持一致。

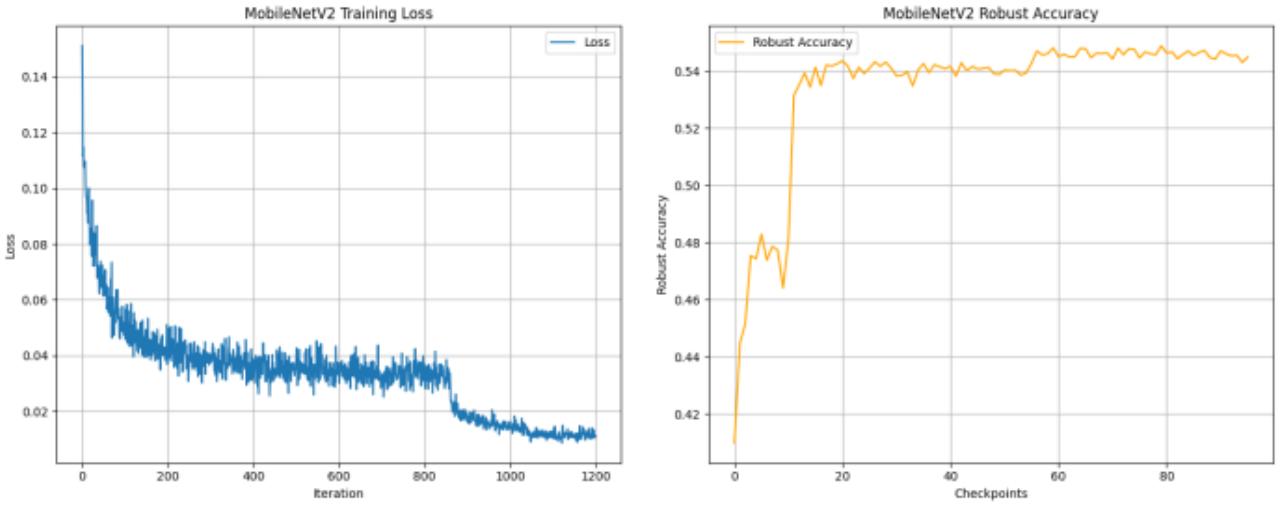

图 2. ResNet-18 训练过程

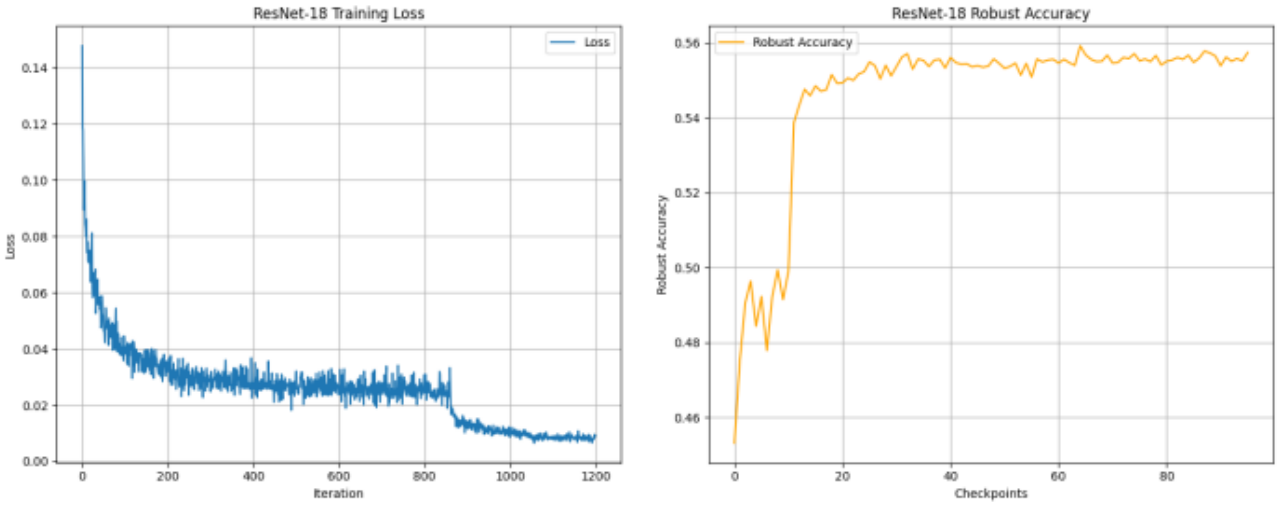

图 3. MobileNetV2 训练过程

Model	Method	Best Checkpoint						Last Checkpoint					
		Clean	FGSM	PGD_{SAT}	PGD_{TRADES}	CW_∞	AA	Clean	FGSM	PGD_{SAT}	PGD_{TRADES}	CW_∞	AA
RN-18	Natural	**94.65%**	19.26%	0.0%	0.0%	0.0%	0.0%	**94.65%**	19.26%	0.0%	0.0%	0.0%	0.0%
	SAT	83.38%	56.41%	49.11%	51.11%	48.67%	45.83%	84.44%	55.37%	46.22%	48.72%	47.14%	43.64%
	TRADES	81.93%	57.49%	52.66%	53.68%	50.58%	49.23%	82.20%	57.86%	52.30%	53.66%	50.69%	49.27%
	ARD	83.93%	59.31%	52.05%	54.20%	51.22%	49.19%	84.23%	59.33%	51.52%	53.74%	51.24%	48.90%
	IAD	83.24%	58.60%	52.21%	54.18%	51.25%	49.10%	83.90%	58.95%	51.35%	53.15%	50.52%	48.48%
	RSLAD	83.38%	**60.01%**	**54.24%**	**55.94%**	**53.30%**	**51.49%**	83.33%	**59.90%**	**54.14%**	**55.61%**	**53.22%**	**51.32%**
MN-V2	Natural	**92.95%**	14.47%	0.0%	0.0%	0.0%	0.0%	**92.78%**	14.59%	0.0%	0.0%	0.0%	0.0%
	SAT	82.48%	56.44%	50.10%	51.74%	49.33%	46.32%	82.89%	56.43%	49.71%	51.48%	49.07%	45.92%
	TRADES	80.57%	56.05%	51.06%	52.36%	49.36%	47.17%	80.57%	56.05%	51.06%	52.36%	49.36%	47.17%
	ARD	83.20%	58.06%	50.86%	52.87%	50.39%	48.34%	83.42%	57.94%	50.63%	52.44%	50.09%	48.01%
	IAD	81.91%	57.00%	51.88%	53.23%	50.45%	48.40%	83.49%	57.44%	49.77%	51.85%	49.41%	46.98%
	RSLAD	83.40%	**59.06%**	**53.16%**	**54.78%**	**51.91%**	**50.17%**	83.11%	**59.08%**	**53.04%**	**54.50%**	**51.60%**	**49.90%**

表 2. 论文实验结果

Model	Method	Clean	FGSM	PGD20	PGD20trades	AA
ResNet-18	论文结果	83.33%	59.90%	54.14%	55.61%	51.32%
	复现结果	**83.54%**	**60.18%**	**54.22%**	**55.73%**	**51.67%**
MobileNetV2	论文结果	83.11%	59.08%	53.04%	54.50%	49.90%
	复现结果	**82.87%**	**59.14%**	**53.39%**	**54.76%**	**50.67%**

表 3. 复现结果（Last Checkpoint）

6 总结与展望

在复现工作中，笔者也尝试性修改过内部最大化损失函数，以此希望提升学生模型的鲁棒性，但可惜未能有多少提升。也尝试过使用具备更强的鲁棒性教师模型，检验是否能产生更强的学生模型，但不非模型越大效果越好，而是需要选择恰当的模型，才能够得到具备最好鲁棒性结果的学生模型。

在本文中，作者提出了一种有效的对抗鲁棒蒸馏方法，通过教师模型教导学生模型，使得小模型也能在对抗样本中表现出很好的鲁棒性。作者以知识蒸馏的角度重新审视最先进的几种对抗训练方法，总结归纳出共同之处，并且发现目前的对抗训练方法仍存在提升的空间，即尚未充分利用鲁棒性软标签和内部最大化损失仍然使用硬标签。通过本次的复现，本文的研究思路清晰明了，也为笔者今后科研方向提供一定的参考价值。

参考文献

[1] Micah Goldblum, Liam Fowl, Soheil Feizi, and Tom Goldstein. Adversarially robust distillation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(04):3996–4003, April 2020.

[2] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network, 2015.

[3] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks, 2019.

[4] Bojia Zi, Shihao Zhao, Xingjun Ma, and Yu-Gang Jiang. Revisiting adversarial robustness distillation: Robust soft labels make student better. In *International Conference on Computer Vision*, 2021.