ADVERSARIAL CONTRASTIVE DISTILLATION WITH ADAPTIVE DENOISING

摘要

对抗性鲁棒蒸馏 (adr) 是一种提高小模型鲁棒性的新方法。与一般的对抗训练不同，其鲁棒的知识迁移不太容易受到模型容量的限制。然而，提供知识鲁棒性的教师模型并不总是做出正确的预测，干扰了学生的鲁棒性表现。此外，在以往的 ARD 方法中，鲁棒性完全来自于一对一的模仿，忽略了实例之间的关系。为此，使用一种新颖的结构化 ARD 方法——对比关系去噪蒸馏 (CRDND)。使用一个自适应补偿模块来对教师的不稳定性进行建模。此外，利用对比关系来探索多个示例之间的隐含鲁棒性知识。在多个攻击基准上的实验结果表明，CRDND 能够有效地迁移鲁棒知识，并取得了最好的性能。

关键词： 对抗防御；鲁棒性蒸馏；

1 引言

深度学习模型在多个领域取得了显著成功，但它们对添加到自然输入中的微小变化很敏感，容易受到攻击 [1–3]。这种脆弱性对于深度学习在关键领域如自动驾驶和金融预测的应用构成了担忧，促使研究人员重新审视模型的鲁棒性 [4]。为了增强模型的对抗鲁棒性，出现了多种策略，尤其是对抗训练（AT） [5]，尽管有效但成本高昂且依赖于模型容量 [6]。特别是在资源受限的环境中，如移动设备和微型机器人上，小型模型的鲁棒性提升尤其具有挑战性，这促使研究者探索了知识蒸馏来改进 AT，即对抗鲁棒性蒸馏（ARD）。

尽管 ARD 的提出减少了 AT 的成本，但仍存在问题。先前的 ARD 方法依赖于预训练的鲁棒模型，而忽略了样本间的相似性，同时教师模型的不稳定预测也限制了学生模型性能的提升。鲁棒知识的传递过于依赖于教师模型，而学生模型的学习潜力并未完全被挖掘，这一点从教师模型的性能限制了学生模型的鲁棒准确性中可见一斑。

论文提出了一种新颖的对抗鲁棒性知识蒸馏方法，命名为对比关系去噪蒸馏（CRDND），通过建立多个样本之间的结构化关系，替代了以往单一的模仿学习方法。这种方法通过探索样本之间的隐含关联，帮助学生模型超越简单的教师模型依赖，实现更好的性能提升。复现本文以便更好地理解鲁棒性蒸馏过程，为后续科研实践打下基础。

2 相关工作

[7] 提出了 ARD 的概念。结果表明，鲁棒模型可以避免昂贵的代价。相反，通过迁移预训练鲁棒模型的知识，小模型可以获得比标准鲁棒训练更高的鲁棒性能。[8] 提出了一种多阶段的策略来进一步提高知识迁移的效率，从而提高学生的鲁棒性。 [6] 认为软目标标签在鲁棒性蒸馏中至关重要，因此他们使用大型鲁棒性模型的完全软目标标签来代替 one-hot 标签，以帮助学生进一步提高鲁棒性。

虽然以往的 ARD 方法可以避免昂贵的 AT 成本，但仍然存在许多问题。一方面，老师的预测并不总是正确的。特别是随着学生的进步，教师对学生生成的对抗样本预测的置信度会逐渐降低 [8]。作为稳健知识的主要来源，这种不稳定的预测限制了学生的表现。 [8] 试图对这种不稳定性进行建模，但实际上，这种不稳定性与教师和学生使用的骨干的能力密切相关。因此，他们的方法并不是所有主干对的通用方法。另一方面，以往的方法都可以归纳为一对一的简单示例模仿学习 [9]。因此，鲁棒知识完全来自预训练的教师，忽略了多个示例之间的隐含相似性。此外，学生模型的学习潜力可能没有得到充分开发，这表现在教师的表现限制了学生的鲁棒准确性。

3 本文方法

3.1 本文方法概述

所提出的对比关系去噪蒸馏的概述如图 1 (a) 所示。遵循之前的 ARD 方法，假设自然样本和固定的、预训练的鲁棒教师模型 (如 WideResNet [10]) 是可用的。然后，目标是训练一个小型学生模型 (例如，MobileNetV2 [11])，同时继承教师的鲁棒性。输入包括自然样本和对抗性样本，以帮助学生处理多个样本场景。为了克服教师预测的不确定性，设计了自适应补偿模块 (Adaptive Compensation Module, ACM) 对不稳定性进行建模。在学生模型的 logit 预测之后添加一个可学习的去噪层，以估计教师答案的正确性。为了提高知识迁移的效率，分别对自然样本和对抗样本进行对比蒸馏，以深入挖掘多样本之间的知识。因此，此方法不完全受限于老师的鲁棒性，取得了很好的效果。

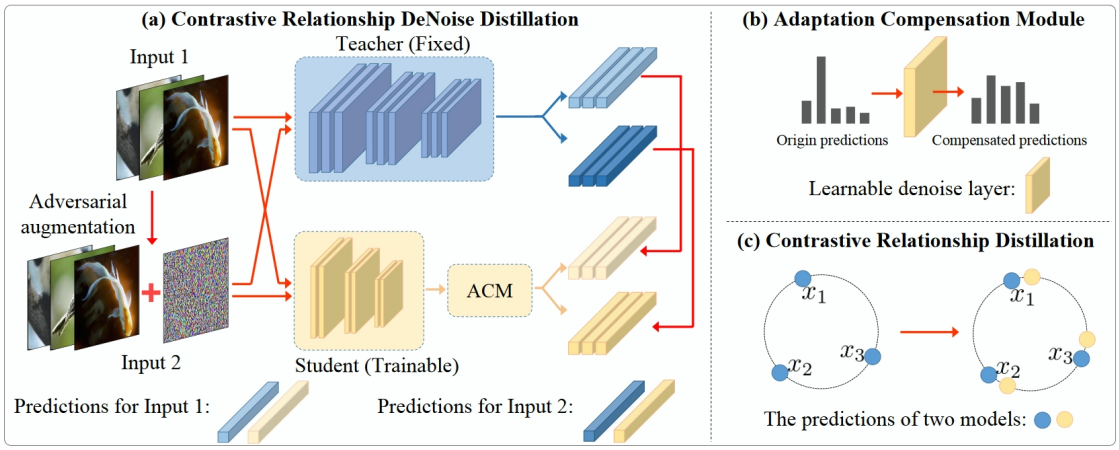

图 1. (a) 对比关系概述去噪蒸馏。(b) 提出的自适应补偿模块 (adaptive compensation module, ACM)。图中列表示预测。(c) 提出的对比关系蒸馏方法。

3.2 自适应补偿模块 ACM

与传统的知识蒸馏 [12] 类似，通过约束教师和学生的预测来迁移知识。我们表示 f_T 和 f_S 为教师和学生模型的 logits 预测，x 为自然样本，x' 为对抗样本。知识转移的过程可以表示为:

$$\mathcal{L} = \sum_{x,x' \in \mathcal{X}} D(f_T(x, x'), f_S(x, x')) \tag{1}$$

其中 D 是距离表示。

然而，如上所述，教师模型的预测并不一定正确。错误的预测往往会导致学生学习 [8] 的错误信息。因此，我们定义了一个可学习的噪声层 $M \in \mathbb{R}^{k \times k}$ 来对教师预测的不稳定性进行建模，如图1(b) 所示。它表示老师预测的正确概率。我们根据对教师精度的估计来调整 M 中的参数。计算了教师模型在当前训练时期关于自然或对抗样本集的准确性。M 的设置规则为: 用当前准确率表示主类权重，对其余类的权重进行平均 (总和为 1)

教师在当前训练时期的真正准确性。M 列可以被认为是一个概率分布，满足 $\sum_{j=1}^{k} M_{ij} = 1$，其中 k 是类别的数量。$M$ 表示为:

$$M = \begin{bmatrix} a_1 & \frac{1-a_2}{k-1} & \cdots & \frac{1-a_k}{k-1} \\ \frac{1-a_1}{k-1} & a_2 & \cdots & \frac{1-a_k}{k-1} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{1-a_1}{k-1} & \frac{1-a_2}{k-1} & \cdots & a_k \end{bmatrix} \tag{2}$$

其中 a_i 表示第 i 类的准确率。M_1, M_2 表示自然场景和对抗性场景的噪声层。那么**??**可以重写为:

$$\mathcal{L}' = \sum_{x,x' \in \mathcal{X}} D(f_T(x, x'), M_1/M_2(f_S(x, x'))). \tag{3}$$

由于该方法只需要估计教师当前的准确率，因此适用于各种教师-学生骨干组合。通过对教师的不稳定性进行建模，该方法还可以克服教师的可靠性在训练阶段逐渐下降的困境。

3.3 对比关系蒸馏

虽然公式(3)可以弥补教师在迁移鲁棒性知识时可能犯的错误，但鲁棒性知识完全取决于教师。为了进一步探索多个示例之间的结构化知识，本文提出对比关系蒸馏方法来替代上述过程，如图 1(c) 所示。重点关注在小批量中教师和学生示例预测的一致性。同时，我们希望分离两个不相对应的预测。在此基础上，构建两种结构化知识来应对自然场景和对抗性场景。首先，对于自然例子 $x = x_1 \ldots x_n$，首先获得鲁棒教师 $f_T(x_i)$ 和噪声补偿学生 $M_1(f_S(x_i))$ 的预测。那么，对比关系可以表示为:

$$L_{\text{nat}}^{x_i} = \frac{\exp(\cos(M_1(f_S(x_i)), f_T(x_i))/T_1)}{\sum_{k=1,k\neq i}^{2N} \exp(\cos(M_1(f_S(x_k)), f_T(x_i))/T_1)} \tag{4}$$

其中 $\tau 1$ 为温度参数，N 为批次大小。接下来，我们可以计算自然样本的关系蒸馏损失:

$$\mathcal{L}_{\text{nat}} = -\frac{1}{N}\sum_{j=1}^{N}\log p_{\text{nat}}^{x_j}. \tag{5}$$

对于对抗性示例 x'，知识表示和转移过程类似于上:

$$\mathcal{L}_{\text{adv}}^{x_i'} = \frac{\exp(\cos(M_2(f_S(x_i')), f_T(x_i'))/T_2)}{\sum_{k=1,k\neq i}^{2N}\exp(\cos(M_2(f_S(x_k')), f_T(x_i'))/T_2)} \tag{6}$$

$$\mathcal{L}_{\text{adv}} = -\frac{1}{N}\sum_{j=1}^{N}\log \mathcal{L}_{\text{adv}}^{x_j'}. \tag{7}$$

最后，我们可以得到总对比关系的降噪蒸馏损失为:

$$\mathcal{L}_{\text{total}} = \lambda\mathcal{L}_{\text{nat}} + (1-\lambda)\mathcal{L}_{\text{adv}}, \tag{8}$$

其中 λ 是损耗权衡参数。与不同，公式(8)模拟了教师和学生之间的一致性以及多个示例之间的差异。来自多个例子的知识是至关重要的，特别是当教师模型不能给出可靠的预测时。

特别值得注意的是，我们的对比学习方法不同于以往的方法。首先，我们的方法不依赖于大的负例集，如大的内存库、大的划分或大的批大小。它也不依赖于额外的归一化和预训练网络更新，计算成本很高。其次，我们的方法不需要仔细设计合适的数据增广运算符。因此，当前方法简单，计算效率高。

4 复现细节

4.1 与已有开源代码对比

本次复现的论文并无代码，因此参考了博客上的代码，但博客代码 ACM 模块存在错误，在参照论文公式和博客经过调试和修改之后，基本复现代码和实验数据和论文基本保持一致。老师模型使用作者公布的预训练模型权重。主要复现工作读取 WideResNet 模型对 cifar10 和 cifar100 的训练权重，根据教师模型训练学生模型 ResNet18 和 MobileNetV2，学生模型中加入了 ACM 模块和对比关系学习模块，再对这两个学生模型分别进行白盒攻击。复现过程中对抗攻击检测额外添加了 CW 攻击测试，目的是检测 CW 攻击下学生模型的鲁棒性。消融实验部分，基于 cifar10 和 cifar100 对学生模型 ResNet18 和 MobileNetV2 的 ACM 模块进行复现，检测其在鲁棒性上的作用。

4.2 复现实验

本次复现实验包括

1. 重新训练四次，包括学生模型 ResNet18 和 MobileNetV2 模型分别在 cifar10 和 cifar100 的蒸馏实验。

2. 分别对学生模型的白盒攻击测试，攻击方法包括 $FGSM$，PGD_{SAT}、PGD_{TRADES}、CW_2（额外添加）、AA。

3. 消除 ACM 模块重新训练学生模型 ResNet18 和 MobileNetV2，再次进行白盒攻击测试验证 ACM 模块对模型的鲁棒性提升效果。

4.3 实验设置和环境

在 CIFAR-10 和 CIFAR-100 这两个常用的对抗鲁棒性测试数据集上对 CRDND 方法进行了评估。基线方法考虑两种 AT 方法:SAT, TRADES, 三种 ARD 方法:ARD, IAD, RSLAD, 以及一种自然训练方法。

老师和学生。为了公平比较，我们选择了 RSLAD [6] 之后相同的教师模型，包括 CIFAR-10 的 WideResNet-34-10 和 CIFAR-100 的 WideResNet-7016。教师模型在整个培训过程中是固定的。此外，我们设置了两个学生骨干，包括 ResNet-18 和 MobileNetV2。

实现细节。所提出的模型在 PyTorch 中实现。我们设定损耗权衡参数为 0.2，温度参数为 0.5。学生通过余弦退火学习率 SGD 优化器进行训练，初始值为 0.1，动量为 0.9，权重衰减为 2e-4。batch size 为 128，总训练步数为 300，与之前的工作相同。对于其他基线方法，我们遵循 RSLAD [6] 的设置。

攻击评估。在多种对抗攻击下评估了该模型:$FGSM$、PGD_{SAT}、PGD_{TRADES} 和 $AutoAttack(AA)$。上述攻击方法与 RSLAD 的设置相同。

采用的是 pytorch 的模型框架，主要的实验环境如下：urllib3、Pillow、numpy、tqdm、tensorboard、torch==1.12.1、torchaudio==0.12.1、torchvision==0.13.1

5 实验结果分析

本论文实验模型在 RSLAD [6] 论文的基础上进行改进，CRDND 数据包括论文数据部分与复现数据部分。

和其他基线方法的鲁棒性性能如表 1 所示。表 1 比较了各种方法中最好的检查点。根据对 PGDT 的鲁棒性，选择了其他训练方法。实验结果表明，CRDND 方法在多个测试集上均取得了较好的鲁棒性。尤其在 FGSM 和 PGD 的评价指标上，ACM 模块和对比学习方法有了很大的提升 (4% - 6%)。一般来说，对抗性鲁棒性的性能与干净条件的性能是一种权衡关系，除非使用了 ground truth 标签 (例如，Nature 和 ARD 方法)。所提出方法在没有任何标签的两种情况下都特别有竞争力，这表明学生通过在多个示例中深入探索额外的知识，提高了整体表现。

表 1. Model performance on CIFAR-10 and CIFAR-100 datasets.

	DataSet	CIFAR-10						CIFAR-100					
Model	Method	Clean	FGSM	PGD_{SAT}	PGD_{TRADES}	CW_2	AA	Clean	FGSM	PGD_{sat}	PGD_{trades}	CW_2	AA
	SAT	84.01%	56.84%	49.11%	50.93%	73.02%	47.12%	56.96%	28.72%	24.33%	25.44%	44.63%	21.94%
	TRADES	81.36%	57.51%	52.66%	54.15%	76.61%	49.69%	55.21%	30.51%	27.85%	28.81%	47.89%	23.63%
	ARD	82.37%	58.36%	51.74%	53.49%	75.85%	49.67%	58.91%	32.52%	28.53%	29.68%	48.34%	25.45%
RN-18	IAD	83.17%	58.96%	52.99%	54.54%	76.63%	50.13%	58.16%	32.71%	28.73%	29.90%	48.19%	25.37%
	RSLAD	83.39%	60.14%	54.26%	55.76%	77.00%	51.91%	58.11%	34.16%	30.73%	31.94%	49.01%	26.67%
	CRDND(原文)	**84.11%**	**64.24%**	**59.91%**	**61.25%**	None	**49.88%**	**59.00%**	**38.02%**	**35.29%**	**36.29%**	None	**27.05%**
	CRDND(复现)	**84.34%**	**65.38%**	**60.61%**	**62.30%**	**83.26%**	**49.44%**	**59.15%**	**38.24%**	**35.51%**	**36.66%**	**59.08%**	**25.93%**
	SAT	81.01%	55.62%	50.66%	52.14%	75.64%	46.85%	58.09%	32.00%	28.72%	29.75%	49.52%	25.14%
	TRADES	79.74%	55.69%	51.55%	52.75%	75.64%	47.12%	56.02%	31.24%	28.81%	29.69%	49.64%	24.21%
	ARD	81.36%	56.17%	51.01%	52.62%	75.53%	48.40%	58.66%	32.86%	29.68%	30.76%	49.23%	26.00%
MN-V2	IAD	81.76%	56.90%	51.70%	53.27%	76.16%	48.33%	57.15%	32.70%	29.93%	31.02%	49.37%	26.09%
	RSLAD	83.17%	58.69%	53.04%	54.51%	77.00%	50.41%	58.56%	33.73%	30.48%	31.52%	49.31%	26.36%
	CRDND	**83.89%**	**65.25%**	**59.93%**	**61.33%**	NONE	**48.79%**	**58.60%**	**38.03%**	**36.05%**	**37.02%**	None	**26.12%**
	CRDND	**84.20%**	**63.45%**	**59.49%**	**60.82%**	**82.04%**	**47.93%**	**58.74%**	**37.95%**	**35.95%**	**36.87%**	**58.37%**	**26.18%**

复现部分包含了自适应补偿模块 (Adaptive Compensation Module, ACM) 的有效性验证，我们将基线设置为 bold(full CRDND) 和 discard ACM，以展示其对结果的影响。表2(1-6) 对比了有无 ACM 对模型鲁棒性的影响 (表中的 w/o 表示无 ACM)。当丢弃 ACM 时，学生对各种攻击指标的鲁棒性都有不同程度的下降。我们认为，下降的原因是缺乏对教师模型预测精度的估计，这表明教师模型给出的频繁错误预测可能会干扰学生的学习。

表 2. Ablation studies on CIFAR-100 dataset (%).

ID	Model	Method	FGSM	PGD_s	PGD_r	AA
1	RN-18	Ours	38.02	35.29	36.29	27.05
2	RN-18	w/o ACM(原文)	**38.02**	**35.14**	**36.11**	**26.29**
3	RN-18	w/o ACM(复现)	**38.00**	**35.12**	**36.12**	**26.27**
4	MN-V2	Ours	38.03	36.05	37.02	26.56
5	MN-V2	w/o ACM(原文)	**37.91**	**35.80**	**36.78**	**26.37**
6	MN-V2	w/o ACM(复现)	**37.94**	**35.80**	**36.77**	**26.42**

6 总结与展望

该文提出了一种新颖的对抗鲁棒性蒸馏方法——对比关系去噪蒸馏 (CRDND)，消除了以往方法对教师模型的完全信任和依赖。首先，设计了即插即用的自适应补偿模块，通过估计教师网络可能的预测误差来校正噪声知识; 其次，提出了一种新的思路，通过挖掘多个样本之间的隐含知识来提高模型的鲁棒性，以应对对抗攻击。在多个攻击基准上的实验结果表明，CRDND 不仅显著提高了学生模型的鲁棒性性能，而且保留了干净样本的性能。

总的来说，该文提出的 CRDND 方法为深度学习模型在对抗环境下的鲁棒性研究提供了新的视角和工具，特别是在资源受限的环境中，这一点具有重要的实际意义。本次复现深入了解了鲁棒性蒸馏方向的相关知识，增强了实践能力，但是并没有过多的创新，未来的研究可以在此基础上进一步提高模型的鲁棒性，或者探索更多结构化知识蒸馏的方法来提升模型性能。

参考文献

[1] Siao Liu, Zhaoyu Chen, Wei Li, Jiwei Zhu, Jiafeng Wang, Wenqiang Zhang, and Zhongxue Gan. Efficient universal shuffle attack for visual object tracking. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2739–2743. IEEE, 2022.

[2] Zhaoyu Chen, Bo Li, Shuang Wu, Jianghe Xu, Shouhong Ding, and Wenqiang Zhang. Shape matters: deformable patch attack. In *European conference on computer vision*, pages 529–548. Springer, 2022.

[3] Jiafeng Wang, Zhaoyu Chen, Kaixun Jiang, Dingkang Yang, Lingyi Hong, Pinxue Guo, Haijing Guo, and Wenqiang Zhang. Boosting the transferability of adversarial attacks with global momentum initialization. *arXiv preprint arXiv:2211.11236*, 2022.

[4] Xiyu Yan, Xuesong Chen, Yong Jiang, Shu-Tao Xia, Yong Zhao, and Feng Zheng. Hijacking tracker: A powerful adversarial attack on visual tracking. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2897–2901. IEEE, 2020.

[5] 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.

[6] 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.

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

[8] 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.

[9] Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3967–3976, 2019.

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

[11] 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.

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