D ^{2} R: 带有协作对抗生成的模型鲁棒性双重正则化损失

Zhenyu Liu[1], Huizhi Liang[1], Rajiv Ranjan[1], Zhanxing Zhu[2], Vaclav Snasel[3], and Varun Ojha[1]

[1] Newcastle University, Newcastle, UK
[2] University of Southampton, Southampton, UK
[3] Technical University of Ostrava, Ostrava, Czech Republic

Abstract. 深度神经网络模型的鲁棒性对于防御模型抵御对抗性攻击至关重要。最近的一些防御方法采用合作学习框架来增强模型的鲁棒性。现有方法的两个主要局限是 (i) 通过损失函数对目标模型的引导不足和 (ii) 非合作的对抗样本生成。因此，我们提出了一种双正则化损失（D ^{2} R 损失）方法和一种对抗训练的合作对抗生成（CAG）策略。D ^{2} R 损失包括两个优化步骤。通过利用在合适的函数空间探索中获得的不同损失函数的优点，针对性地增强目标模型分布的对抗分布和干净分布的优化，提高目标模型的鲁棒性。CAG 通过引导模型和目标模型之间的基于梯度的合作生成对抗样本。我们在包括 CIFAR-10、CIFAR-100、Tiny ImageNet 在内的三个基准数据库以及两个流行目标模型 WideResNet34-10 和 PreActResNet18 上进行了广泛实验。我们的结果表明，结合 CAG 的 D ^{2} R 损失能够生成高度鲁棒的模型。我们的代码可在 https://github.com/lusti-Yu/D2R.git 获得。

1 引言

深度神经网络（DNNs）在解决现实世界中的问题中至关重要。然而，DNNs 易受到对抗性攻击的影响 [14, 19] 。研究人员一直积极研究防御策略，以提高 DNN 模型对抗对抗性攻击的鲁棒性。最有效和众所周知的防御方法之一是对抗性训练（AT） [11] 。多项研究表明，对大型模型进行对抗性训练比小型模型提供更大的鲁棒性（由于可学习参数较多） [12] 。同样，对抗性蒸馏（AD）方法使用预训练的教师模型来增强学生模型（目标模型）的鲁棒性 [10, 2] 。AD 方法的替代方案是一组对抗性防御策略，这些策略涉及使用小模型作为引导来支持目标模型。这些小型引导模型和目标模型被同时训练以增强目标模型的鲁棒性。LBGAT 是引导与目标框架的一个例子 [1] 。使用这个小型可学习引导模型来支持目标模型的优势是 (a) 它能够自适应地参与对抗性训练过程，(b) 不需要一个大型预训练模型作为引导，以及 (c) 可学习的引导模型可以支持任何尺寸目标模型的对抗性训练。

然而，之前的工作存在两个主要限制。(i) 基于更广泛的损失函数空间探索的约束，即，大多数工作使用平方误差损失通过利用一个可学习的指导模型（例如，[1] [8] ）的特性来实现对抗训练。这使得对抗训练依赖于选定的损失函数，并阻碍其达到最佳性能。也就是说，损失函数的非最优选择导致对目标模型指导不足。(ii) 基于对抗样本生成的约束，即，大多数防御方法（如 TRADES [18] 和 PGD-AT [11] ）使用单一模型和独立的对抗样本生成技术。这种非协作的对

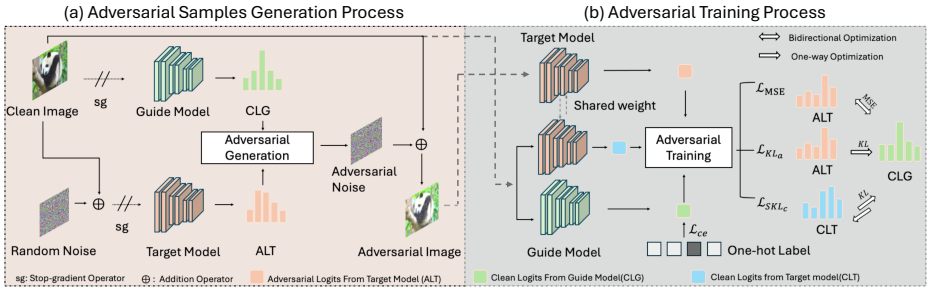


Fig. 1: 我们 D ^{2} R-CAG 框架的概述。(a) 提出的协作对抗生成过程，其中引导模型和目标模型创建扰动。这些扰动应用于图像以生成对抗性图像样本。(b) D ^{2} R 损失。$\mathcal{L}ce$ 表示引导模型的干净训练损失，$\mathcal{L}_{MSE}$ 表示目标模型和引导模型之间的均方误差，$\mathcal{L}_{KL_a}$ 是通过 KL 散度优化对抗性分布，$\mathcal{L}_{SKL_c}$ 是使用对称 KL 散度函数优化干净分布。在整个对抗训练过程中，引导和目标模型使用同步输入进行训练。

抗生成阻碍了基于可学习的指导-目标架构的对抗训练，因为它忽略了指导模型和目标模型之间的协作，从而无法达到最佳性能。

我们提出了一种双重正则化损失（D ^{2} R 损失）方法，以减轻在引导-目标框架中选择损失函数的限制（见图 1 (b)，显示对抗训练过程）。D ^{2} R 损失（图 1 (b)）利用不同损失函数的优势，精准聚焦于引导模型和目标模型的分类分布，以增强目标模型的鲁棒性。我们将 D ^{2} R 损失过程分解为两个步骤：(i) 对抗分布优化和 (ii) 干净分布优化。具体而言，a) 使用对抗性的 KL 散度（$\mathcal{L}_{KL_a}$）来实现引导模型的干净分类输出分布（logits）和目标模型的对抗分类输出分布（logits）之间的精确对齐；以及 b) 使用干净的 KL 散度（$\mathcal{L}_{SKL_a}$）有助于引导模型的干净分类输出分布（logits）和目标模型的干净分类输出分布（logits）之间的精确对齐。我们利用 KL 散度的不对称性，轻微且动态地估计引导模型和目标模型之间的干净分布差距，从而增强模型的泛化能力，进而提高鲁棒性。

我们提出了一种可学习的协同对抗生成（CAG）策略，将对抗样本生成过程整合到整体对抗训练框架中（参见图 1 （a），显示对抗样本生成过程），而不是像以往工作中那样让对抗样本生成在独立状态下进行。CAG 策略基于包含指引模型和目标模型分布的全局性能。在这种对抗样本生成过程中，我们还利用了可学习的指导模型在其对干净样本的正确预测上的高置信度。随着训练的进行，指导模型和目标模型协同生成对抗样本。这种协作方法旨在生成自适应的对抗样本，从而增强模型的鲁棒性。总之，这项工作的主要贡献如下：

- 我们提出了一种双重正则化损失（D ^{2} R Loss），旨在逐步增强目标模型的鲁棒性。
- 我们提出了一种协作对抗生成（CAG）方法，该方法可动态创建用于后续对抗训练的对抗样本。
- 我们表明，我们的方法（D ^{2} R Loss with CAG）在大量实验中优于其他 AT 方法。

2 相关工作

对抗训练是最广泛使用的防御方法。例如，Wu 等人提出了一种有效的对抗训练方法，即对抗权扰动（AWP），以提高 DNN 的鲁棒性。类似地，AT 方法 RAT 通过向 DNN 权重添加随机噪声来展示模型的鲁棒性，而误分类感知对抗训练（MART）分析误分类样本以提高 DNN 模型的鲁棒性。LAS-AT 使用单一策略模型将攻击注入目标模型，而不是像传统 AT 中广泛使用的那样手动设计攻击。TRADES 优化了准确性和鲁棒性之间的权衡，通过 KL 散度损失生成对抗样本进行训练。

依赖单一模型生成对抗样本在增强目标深度神经网络的鲁棒性方面起着重要作用。然而，使用单一模型生成对抗样本在实现各种场景的最佳性能方面存在不足。因此，我们的工作重点是使用一个可学习的引导模型来自适应地与目标模型协作，以生成对抗样本。

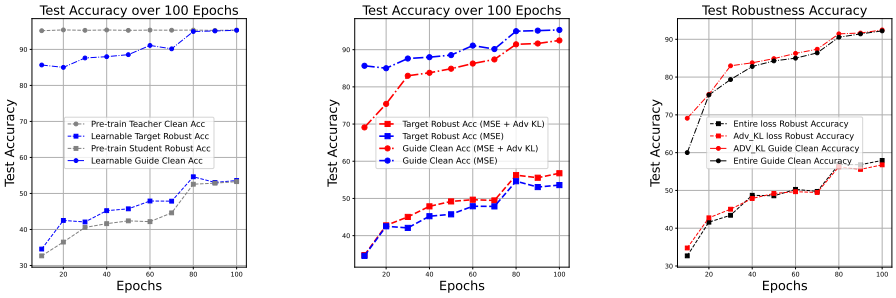
我们的防御方法与现有的知识蒸馏方法显著不同。传统的知识蒸馏通常使用一个预训练的引导模型作为静态的真实概率分布，并仅使用 KL 散度来优化目标模型的参数。此外，关注使用 KL 散度的方法，如 MTARD [22]，倾向于仅检查干净教师模型和鲁棒教师模型之间的知识尺度差距。然而，他们并不专注于最小化这些分布差异，并且他们的教师模型在训练过程中是不可学习的。相比之下，我们的研究独特地专注于优化可学习引导模型和目标模型之间的干净概率分布，提供了一种新颖的视角来指导目标对抗性蒸馏/对抗性训练框架。此外，我们使用小型、未训练的模型作为引导模型，并同时训练引导模型和目标模型进行知识蒸馏。在此过程中，我们使用我们的双重正则化损失来增强损失函数，而不是其他方法，如 LBGAT [1]，这些方法仅使用 MSE 来优化引导和目标错误之间的差异。

3 提出的方法

3.1 设计原则

我们考虑一个干净数据集为 $D = \{(x_i, y_i)\}$，其中 $x_i \in \mathbb{R}^d$ 是一个 d 维的输入，$y \in [k]$ 是一个类别标签。我们有一个分类器 $f_\theta(x)$，它将输入映射到标签空间为 $f_\theta : x \to y$，使用一些可学习的参数 $\theta \in \mathbb{R}^p$，也称为模型参数。当对于一个干净的输入 x 存在一个对抗性输入 $x^{'} = x + \delta$ 时，该模型被攻击，从而模型性能下降，其中 $\delta \in \mathbb{R}^d$ 是对抗性扰动，并根据 PGD-AT [11]，δ 被限制在 l_∞ 内。对抗蒸馏（AD） [10] 已经证明了其作为一种防御方法的有效性，使模型对这种攻击具有鲁棒性。通常，AD 涉及一个预训练的指导模型。尽管对抗蒸馏在各种研究中显示了令人印象深刻的性能（例如，[4]），但基于较小模型的可学习框架也能取得显著的结果（例如， [1, 5]）。因此，我们提出了一种基于可学习架构的新对抗防御方法。具体而言，传统的对抗蒸馏方法使用静态的、预训练的模型将知识转移到目标模型。然而，预训练模型具有有限的适应性。受到可学习指导模型 [1] 和预训练教师模型 [10] 的启发，我们的目标是通过探索预训练方法和可学习方法之间的差距，进一步研究可学习指导模型的鲁棒性优势。

我们对可学习指导模型（ResNet-18）和预训练（ResNet-18）指导模型进行了公平比较，以验证我们的假设。实验表明，可学习指导模型更有效地增强了鲁棒性（参见图 2 (a)）。这表明作为教师的可学习指导模型比固定预训练模型作为教师更能适应对抗训练。因此，我们在可学习指导模型上进行了实验，同时使



(a) (b) (c)

Fig. 2: 在 CIFAR-10 上的 PGD-20 攻击下模型鲁棒性比较。(a) 预训练的 ResNet-18（灰色）与可学习的 ResNet-18（蓝色）。(b) 使用我们的方法引导模型的干净准确率和目标模型的鲁棒性（红色）与 LBGAT（蓝色）的对比。(c) 干净对称损失的效果：没有（红色）与有（黑色）的对比。

用基线方法 [1]，该方法使用均方误差（MSE）作为对抗训练损失。相比之下，图 2 中的另一个指导模型使用了 MSE 和 KL-divergence 的组合损失作为对抗训练损失。实验显示，相较于仅使用 MSE，组合损失提供了更好的鲁棒性（参见图 2 (b)）。同样，图 2 (c) 显示了在干净输入上指导和目标模型输出分布有助于实现更好的性能。尽管可学习架构提供了鲁棒性方面的改进，由于指导和目标模型之间的复杂交互，设计一个有效的指导策略是具有挑战性的，这需要目标模型更多地全面参与。这一观察促使我们提出一个更加精细和综合的过程。

我们的工作进行了一项功能空间探测 [7]，旨在增强鲁棒性准确率。当涉及多个 DNN 时，损失函数之间的交互损失地貌尚未被很好理解。然而，已知损失函数的性质：CE 衡量预测分数的相似性（全球性），MSE 衡量两组 logits 之间的差异（局部性），而 KL 散度衡量两个概率分布之间的差异。即，CE 在干净分类任务中展示了有效性，MSE 损失常用于可学习指导-目标架构，而 KL 是对抗蒸馏的热门选择。同时，CE 损失的脆弱性已被充分理解 [9]。多重损失框架利用了一组不同损失函数 [17] 之间的正则化。在我们的框架中，由于两个 DNN 在两个不同任务间相互作用，损失函数涉及整体损失正则化的任务特定组件。

整个双重正则化损失 D^2R（图 1 (b)）可以分为两个部分。首先，对抗损失通过采用 KL 散度损失来增强目标模型的鲁棒性。其次，结合指南和目标模型之间的干净分布优化可以有效提高清晰度和鲁棒性。此外，协作对抗生成策略（图 1 (a)）可以进一步提高目标模型的鲁棒性。

虽然可学习架构的基本概念已证明其有效性，但 MSE 损失函数仍然遇到限制其鲁棒性的精度问题。此外，我们假设如果引导模型能够更准确地学习目标模型的对抗分布，就会提供更加同步和有效的指导。

上述观察和假设促使我们提出一种新方法，该方法更准确地关注目标模型的分布，超越使用 MSE 函数来近似全局平方差分布。我们建议根据引导模型的清洁分布来限制目标模型的对抗性分布。我们的设计结合了对 MSE 和 KL 散度函数的操作，而不只是单独使用 MSE，因为 KL 散度对微小的概率变化非常敏感。从图 2 (b) 中，我们观察到与单独依赖 MSE 相比，我们的方法在增强目标模型

的鲁棒性方面更有效。此外，我们提出的对抗损失使目标模型能够适应引导模型的清洁分布。该方法使目标模型的边界能够更准确地与引导模型的清洁状态对齐。结合对抗优化后，引导模型的清洁精度和目标模型的鲁棒性均有所提高，如图 2 (c) 所示。因此，通过结合 KL 散度和 MSE 损失，增强鲁棒性精度的目标可以得到更好的实现。

根据前面的分析和函数属性的视角，KL 散度捕捉概率分布的微妙差异和结构特征，而 MSE 强调整体的平均误差。这种双重结合利用 KL 散度来解决分布的差异，并使用 MSE 来量化整体偏差以提高鲁棒性。因此，我们有一个有效的对抗性分布指导（ADG）损失，如下所示：

$$\begin{aligned}\mathcal{L}_{\mathrm{ADG}}(x,y) = \min_{\theta_g,\theta_t} \mathbb{E}_{(x,y)\in D}\big\{&\mathcal{L}_{\mathrm{CE}}(\theta_g,x,y)\\ &+\mathcal{L}_{\mathrm{MSE}}(f_g(x),f_t(x^{'}))+\alpha\mathcal{L}_{\mathrm{KL}}(f_g(x)\parallel f_t(x^{'}))\big\},\end{aligned} \tag{1}$$

其中 θ_g 和 θ_t 是指导和目标模型的参数；$\mathbb{E}_{(x,y)\in D}$ 是所有输入输出对 $(x,y)\in D$ 上经验风险的期望；$\mathcal{L}_{\mathrm{CE}}$, $\mathcal{L}_{\mathrm{MSE}}$ 和 $\mathcal{L}_{\mathrm{KL}}$ 分别是交叉熵、均方误差和 KL 散度的损失函数。项 $f_g(x)$ 表示指导模型的干净分布，而 $f_t(x^{'})$ 表示目标模型的对抗性分布。$\alpha\geq 0$ 是用户定义的超参数，用于控制等式 (1) 中 KL 散度的强度。

为进一步增强模型的鲁棒性，我们强调对干净输出的近似。在两个可共同训练的模型的训练过程中，我们旨在提出一个针对干净输出的最优近似策略，以确保目标模型的鲁棒性提升。虽然之前的研究中使用了 KL 散度进行模型之间的互学习，证明了通过改善蒸馏任务中的泛化能力可收敛于一个更鲁棒的极小值，但其在实现对抗性训练中的应用尚未得到充分探索。相比之下，我们的方法旨在通过动态优化引导或目标模型以实现更好的泛化来提高目标模型的鲁棒性。我们采用对称 KL 项之间的差异来优化引导模型的概率分布。具体来说，我们的方法通过利用 KL 函数中真实和目标概率分布的位置差异来近似不同 KL 散度之间的差异。我们结合了 KL 不对称性中的细微差别。根据对抗性优化中的方程，在忠实重构 D ^{2} R 损失的过程中，整个过程被推导为：其中 $\lambda\in\mathbb{R}$ ，$\alpha\geq 0$ ，$\beta\in\mathbb{R}$ 是用于控制不同损失函数贡献的用户定义超参数。结合方程有效地解决了引导模型和目标模型之间干净分布的巨大差距。具体来说，基于这种整合，该方法可以有效地减轻目标模型因干净样本的泛化改进而导致的错误分类（如图所示），进一步增强了目标模型的鲁棒性，这在图中得到验证。

备注。我们将 $\mathcal{L}_{\mathrm{KL}}(f_t(x)\parallel f_g(x))$ 表示为一个 KL 散度，用于量化引导模型的干净分布在多大程度上近似目标模型的干净分布。相反，$\mathcal{L}_{\mathrm{KL}}(f_g(x)\parallel f_t(x))$ 量化目标模型的干净分布在多大程度上近似引导模型的干净分布。因此，我们有

$$\mathcal{L}_{\mathrm{KL}}(f_t(x)\parallel f_g(x))\neq\mathcal{L}_{\mathrm{KL}}(f_g(x)\parallel f_t(x)) \tag{2}$$

。

在我们的函数空间探索中，对于在纯净分布下的引导-目标模型，我们有两种不同的优化场景。当值为正时，我们仅优化目标模型的概率分布 $f_t(x)$ 。相反，当值为负时，我们专注于优化引导模型的概率分布 $f_g(x)$ 。这种动态方法允许我们根据各自的状态灵活调整 $f_g(x)$ 或 $f_t(x)$ ，保持引导和目标纯净分布之间的一致性。从经验上看，我们观察到该值在正负之间略有波动，双方都不占据优势。因此，这种在引导和目标模型之间调整优化重点的策略被经验认为比单一固定优化更有效。

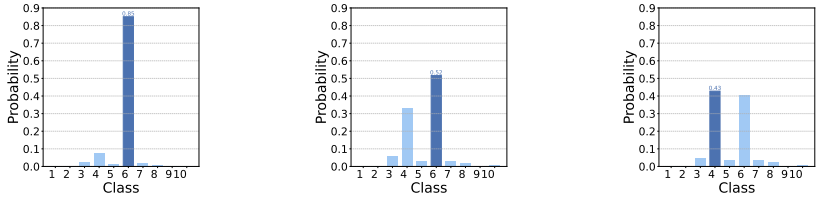


Fig. 3: 一个展示对称 KL 有效性的示例。（左）引导模型输出的概率分布被正确分类为第 6 类（与真实标签相同）。（中）在应用对称清洁分布后，目标模型的概率分布被分类为第 6 类。（右）在未应用对称清洁分布的情况下，目标模型的概率分布被错误分类为第 4 类。

在此，我们利用 KL 散度的不对称性，通过缩小引导模型和目标模型之间的干净概率分布的差距，以最小化知识的泄漏，从而可以提高目标模型的干净和鲁棒性准确率。因此，我们观察到，当对抗性输出知识充分对齐时，干净样本上的增量泛化进一步有助于增强鲁棒性。

协作对抗生成 对抗生成在稳健性中起着决定性作用 [5] 。在文献中，像 TRADES [18] 这样的方法通过约束对抗样本使用来自单一模型的 KL 散度进行生成。然而，这种方法在缓解对抗稳健性挑战方面显得不足。

由于对抗生成与对抗梯度的期望纠缠在一起，直接基于复杂框架解决这样的优化问题具有挑战性。相反，我们提出了一种交替对抗生成方法，模型引导和目标模型协同参与该过程。对于我们的对抗样本生成，我们通过采用基于协作梯度的方法，结合给定的指导模型和目标模型，来生成对抗扰动。也就是说，对抗扰动是通过整个框架迭代生成的。我们采用一个指导模型与目标模型一起自动生成对抗样本 $x^{''}$ 。我们的方法的关键方面如下：

$$x^{''} = \Pi\big(\eta\,\mathrm{sign}(\nabla_{x'}\mathcal{L}_{\mathrm{KL}}(f_t(x^{'}) \parallel f_g(x))) + x^{'}\big), \tag{3}$$

其中 $\mathcal{L}_{\mathrm{KL}}$ 表示我们实验中使用的 KL 散度损失，$x^{'} = x + \delta$ 是随机选择的对抗样本，$f_t(x^{'})$ 表示当以 $x^{'}$ 作为输入时从目标模型得到的对抗逻辑值。我们提出的生成损失函数衡量的是对抗目标逻辑和干净引导逻辑之间的差异。值得注意的是，我们的协作对抗生成（CAG）方法整合了双重正则化损失（D ^{2} R Loss）方法。这种整合是关键的，因为生成对抗样本要求目标模型和指导模型的概率分布保持在一个指定范围内。

我们在 CIFAR-10、CIFAR-100 和 Tiny ImageNet 上进行了大量实验，以评估我们方法在各种白盒攻击下的鲁棒和干净准确性。

对抗防御的评估。为了验证我们提出方法的有效性，我们将我们的防御方法（D ^{2} R 损失不使用 CAG 和 D ^{2} R 损失使用 CAG）与文献中的几种防御方法及其扩展版本在稳健性和干净准确性方面进行了比较。评估的方法包括一系列防御方法，包括 1) PGD-AT，2) TRADES，3) SAT，4) MART，5) FAT，6) GAIRAT，7) AWP，8) LBGAT，9) LAS-AT，以及 10) RAT。

模型设置和实施细节。为了确保结果的可比性和实验效率。根据 [1] ，我们采用 WideResNet34-10 作为目标网络的主干和相应的指导模型（ResNet-18 或

PreActResNet18)。我们将扰动幅度设置为 0.031，扰动步长为 0.007，迭代次数为 10，学习率为 0.1。批量大小设置为 128。

在 CIFAR-10 和 CIFAR-100 上的比较。CIFAR-10 和 CIFAR-100 的结果显示在表格 1 中。我们注意到当与 RAT 和 LAS-AT 分别结合时，AWP 方法之间存在差异。因此，为了进行真正公平的比较，我们避免加入额外的技术（即 AWP），除非明确需要（例如 TRADES）。表格 1 展示了与 CIFAR-10 数据集的比较结果。对于鲁棒性精度，D ^{2} R 在所有攻击情境下表现优于 LBGAT 方法。此外，我们的 D ^{2} R-CAG 方法在 PGD-50 攻击和 AA 攻击下分别提升了 LBGAT 的性能约 2.3 % 和 1.81 % 。此外，所提出的 D ^{2} R-CAG 在 C & W 和 AA 攻击情境下实现了最佳的鲁棒性表现。对于 CIFAR-100，所提出的 D ^{2} R-CAG 在所有攻击情境下实现了最佳的鲁棒性表现。详细而言，我们的 D ^{2} R-CAG 在干净精度和 PGD-50 攻击精度上分别比 LBGAT 提高了 1.08 % 和 2.47 % 。值得注意的是，对于我们的 D ^{2} R-CAG，我们将这些改进归因于我们使用了自动生成的攻击策略，而不是依赖于 TRADES。

在 Tiny ImageNet 上的比较。遵循 [5] 设置，我们将 PreActResNet18 [3] 设为对象模型进行 Tiny ImageNet 的评估。结果如表 1 所示。而且，我们的 D ^{2} R-CAG 在所有攻击场景以及干净准确性方面均优于 LAS-AT。此外，我们进一步使用相同的架构、超参数和纪元提供 LBGAT 方法与 D ^{2} R-CAG 的简单比较。结果表明我们的方法显著优于 LBGAT，表现更佳。具体来说，我们的 D ^{2} R-CAG 超过了 LBGAT 1.19 % AA 攻击准确性。

消融研究我们对以下三个参数进行了消融实验：D ^{2} R 损失中的两个组成部分，即对抗性分布优化损失和干净分布优化，以及指导模型训练中的交叉熵（CE）损失。为了验证分布优化中各组件的有效性，我们在 CIFAR-10 和 CIFAR-100 上进行了消融实验。对于 CIFAR-10，我们训练了具有不同对抗性分布和干净分布优化参数的模型。然后对训练的模型进行了多种对抗性攻击方法的测试。CIFAR-10 的结果表明，每个损失函数在鲁棒性方面都达到了最佳性能。对于 CIFAR-100，我们也尝试了交叉熵（CE）损失的不同参数。结果表明，这三个组件是兼容的，它们的联合使用共同增强了模型的鲁棒性。

超参数的敏感性在我们提出的算法中，正则化超参数 λ ，α 和 β 是至关重要的。我们观察到，随着正则化参数 α 的增加，稳健性和清晰度的准确性都在提高。准确性对正则化超参数非常敏感。考虑到稳健性准确性和清晰度准确性之间的权衡，不难发现，$\alpha = 30$ 和 $\beta = 20$ 适合用于外部优化。在 CIFAR-10 中的参数设置中，我们使用 $\lambda = 1.0$ 。对于 CIFAR-100 和 Tiny ImageNet 的交叉熵项优化，使用 $\lambda = 0.1$ 。

4 结论

我们引入了 D ^{2} R 损失函数，该损失函数通过三种方法逐步提高模型的鲁棒性。可以从各种损失函数对分布的影响角度理解我们的方法。此外，基于全局分布生成对抗样本也可以增强模型的鲁棒性。最后，在 CIFAR-10、CIFAR-100 和完整的 Tiny ImageNet 数据集上进行的大量实验证明了我们方法的有效性。

References

1. Cui, J., Liu, S., Wang, L., Jia, J.: Learnable boundary guided adversarial training. In: Proceedings of the IEEE/CVF International Conference on Computer Vision.

Table 1: 在 CIFAR-10 和 CIFAR-100 上使用 WRN34-10 的鲁棒性 (%)。TinyImgNet [1] 使用与 LAS-AT 中相同的 PreActResNet18。TinyImgNet [2] 对完整的 Tiny ImageNet 使用 WRN34-10 进行训练，并使用 ResNet18 作为 LBGAT 和我们方法的指导。最佳结果以粗体显示。

Dataset	Method	Clean	PGD-50	C & W	AA
CIFAR-10	PGD-AT [11]	85.17	54.88	53.91	51.69
	TRADES [18]	85.72	55.9	53.87	53.40
	MART [15]	84.17	58.06	54.58	51.10
	FAT [20]	87.97	48.79	48.65	47.48
	GAIRAT [21]	86.30	58.74	45.57	40.30
	AWP [16]	85.57	57.92	56.03	53.90
	LAS-AT [5]	86.23	56.12	55.73	53.58
	RAT(TRADES) [6]	85.98	-	56.13	54.20
	LBGAT [1] (baseline)	88.22	54.30	54.29	52.23
	D [2] R-CAG (ours)	85.68	56.73	56.66	54.65
CIFAR-100	PGD-AT [11]	60.89	31.45	30.1	27.86
	TRADES [18]	58.61	28.56	27.05	25.94
	SAT [13]	62.82	26.76	27.32	24.57
	AWP [16]	60.38	33.65	31.12	28.86
	LAS-AT [5]	61.80	32.54	31.12	29.03
	RAT(TRADES) [6]	63.01	-	29.44	28.10
	LBGAT [1] (baseline)	60.64	34.62	30.65	29.33
	D [2] R-CAG (ours)	61.72	35.01	31.59	29.61
TinyImageNet [1]	PGD-AT [11]	43.98	19.98	17.6	13.78
	TRADES [18]	39.16	15.74	12.92	12.32
	LAS-AT [5]	44.86	22.16	18.54	16.74
	D [2] R-CAG (ours)	49.42	22.98	19.29	16.96
TinyImageNet [2]	LBGAT [1] (baseline)	49.11	24.41	20.22	18.32
	D [2] R-CAG (ours)	49.47	24.96	21.10	19.51

Table 2: 消融研究的表现。在 CIFAR-10 上测试模型 WRN34-10 的鲁棒性 (%)。最佳鲁棒性用粗体表示。

Method	Clean	PGD-10	PGD-20	PGD-50	C & W	AA
D [2] R ($\alpha = 0$, $\beta = 1$)	88.35	55.70	54.28	53.85	53.82	52.80
D [2] R ($\alpha = 1$, $\beta = 0$)	88.14	55.62	54.17	53.75	53.57	51.69
D [2] R ($\alpha = 20$, $\beta = 1$)	86.48	57.33	56.23	55.74	55.02	**53**.52
D [2] R ($\alpha = 30$, $\beta = 1$)	85.87	57.85	56.83	56.50	55.62	53.87
D [2] R ($\alpha = 30$, $\beta = 20$)	86.00	58.17	56.88	56.60	55.69	54.04
D [2] R-CAG ($\alpha = 30$, $\beta = 20$)	85.68	58.50	57.22	56.73	56.66	54.65

pp. 15721–15730 (2021)
2. Goldblum, M., Fowl, L., Feizi, S., Goldstein, T.: Adversarially robust distillation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp.

3996–4003 (2020)
3. He, K., Zhang, X., Ren, S., Sun, J.: Identity mappings in deep residual networks. In: 14th European Conference Computer Vision. pp. 630–645 (2016)
4. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. arXiv:1503.02531 (2015)
5. Jia, X., Zhang, Y., Wu, B., Ma, K., Wang, J., Cao, X.: Las-at: adversarial training with learnable attack strategy. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13398–13408 (2022)
6. Jin, G., Yi, X., Wu, D., Mu, R., Huang, X.: Randomized adversarial training via taylor expansion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16447–16457 (2023)
7. Lebedev, V.I.: An Introduction to Functional Analysis in Computational Mathematics: An Introduction. Springer (1996)
8. Liu, Z., Duan, H., Liang, H., Long, Y., Snasel, V., Nicosia, G., Ranjan, R., Ojha, V.: Dynamic label adversarial training for deep learning robustness against adversarial attacks. International Conference on Neural Information Processing (2024)
9. Pang, T., Xu, K., Dong, Y., Du, C., Chen, N., Zhu, J.: Rethinking softmax cross-entropy loss for adversarial robustness. In: International Conference on Learning Representations (2020)
10. Papernot, N., McDaniel, P., Wu, X., Jha, S., Swami, A.: Distillation as a defense to adversarial perturbations against deep neural networks. In: IEEE Symposium on Security and Privacy. pp. 582–597 (2016)
11. Rice, L., Wong, E., Kolter, Z.: Overfitting in adversarially robust deep learning. In: International Conference on Machine Learning. pp. 8093–8104 (2020)
12. Schmidt, L., Santurkar, S., Tsipras, D., Talwar, K., Madry, A.: Adversarially robust generalization requires more data. In: Advances in Neural Information Processing Systems (2018)
13. Sitawarin, C., Chakraborty, S., Wagner, D.: Sat: Improving adversarial training via curriculum-based loss smoothing. In: Proceedings of the 14th ACM Workshop on Artificial Intelligence and Security. pp. 25–36 (2021)
14. Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., Fergus, R.: Intriguing properties of neural networks. In: International Conference on Learning Representations (2014)
15. Wang, Y., Zou, D., Yi, J., Bailey, J., Ma, X., Gu, Q.: Improving adversarial robustness requires revisiting misclassified examples. In: International Conference on Learning Representations (2019)
16. Wu, D., Xia, S.T., Wang, Y.: Adversarial weight perturbation helps robust generalization. Advances in Neural Information Processing Systems **33**, 2958–2969 (2020)
17. Xu, C., Lu, C., Liang, X., Gao, J., Zheng, W., Wang, T., Yan, S.: Multi-loss regularized deep neural network. IEEE Transactions on Circuits and Systems for Video Technology **26**(12), 2273–2283 (2015)
18. Zhang, H., Yu, Y., Jiao, J., Xing, E., El Ghaoui, L., Jordan, M.: Theoretically principled trade-off between robustness and accuracy. In: International Conference on Machine Learning. pp. 7472–7482 (2019)
19. Zhang, J., Huang, Y., Xu, Z., Wu, W., Lyu, M.R.: Improving the adversarial transferability of vision transformers with virtual dense connection. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 7133–7141 (2024)
20. Zhang, J., Xu, X., Han, B., Niu, G., Cui, L., Sugiyama, M., Kankanhalli, M.: Attacks which do not kill training make adversarial learning stronger. In: International Conference on Machine Learning. pp. 11278–11287 (2020)

21. Zhang, J., Zhu, J., Niu, G., Han, B., Sugiyama, M., Kankanhalli, M.: Geometry-aware instance-reweighted adversarial training. In: International Conference on Learning Representations (2020)
22. Zhao, S., Wang, X., Wei, X.: Mitigating accuracy-robustness trade-off via balanced multi-teacher adversarial distillation. IEEE Transactions on Pattern Analysis and Machine Intelligence (2023)