基于随机平滑的通用黑盒认证防御

李 瞧[1] 陈 晶[1),2)] 张子君[1)] 何 琨[1)] 杜瑞颖[1),3)] 汪欣欣[1)]

[1)](空天信息安全与可信计算教育部重点实验室 武汉大学国家网络安全学院 武汉 430079)

[2)](武汉大学日照信息研究院 山东 日照 276800)

[3)](地球空间信息技术协同创新中心 武汉 430079)

摘 要 近年来,基于深度神经网络(DNNs)的图像分类模型在人脸识别、自动驾驶等关键领域得到了广泛应用,并展现出卓越的性能.然而,深度神经网络容易受到对抗样本攻击,从而导致模型错误分类.为此,提升模型自身的鲁棒性已成为一个主要的研究方向.目前大部分的防御方法,特别是经验防御方法,都基于白盒假设,即防御者拥有模型的详细信息,如模型架构和参数等.然而,模型所有者基于隐私保护的考虑不愿意共享模型信息.即使现有的黑盒假设的防御方法,也无法防御所有范数扰动的攻击,缺乏通用性.因此,本文提出了一种适用于黑盒模型的通用认证防御方法.具体而言,本文首先设计了一个基于查询的无数据替代模型生成方案,在无需模型的训练数据与结构等先验知识的情况下,利用查询和零阶优化生成高质量的替代模型,将认证防御场景转化为白盒,确保模型的隐私安全.其次,本文提出了基于白盒替代模型的随机平滑和噪声选择方法,构建了一个能够抵御任意范数扰动攻击的通用认证防御方案.本文通过分析比较原模型和替代模型在白盒认证防御上的性能,确保了替代模型的有效性.相较于现有方法,本文提出的通用黑盒认证防御方案在 CIFAR10 数据集上的效果取得了显著的提升.实验结果表明,本文方案可以保持与白盒认证防御方法相似的效果.与之前基于黑盒的认证防御方法相比,本文方案在实现了所有 L_p 的认证防御的同时,认证准确率提升了 20%以上.此外,本文方案还能有效保护原始模型的隐私,与原始模型相比,本文方案使成员推理攻击的成功率下降了 5.48%.

关键词 深度神经网络;认证防御;随机平滑;黑盒模型;替代模型

中图法分类号 TP391 **DOI 号** 10.11897/SP.J.1016.2024.00690

Universal Certified Defense for Black-Box Models Based on Random Smoothing

LI Qiao[1)] CHEN Jing[1),2)] ZHANG Zi-Jun[1)] HE Kun[1)] DU Rui-Ying[1),3)] WANG Xin-Xin[1)]

[1)](*Key Laboratory of Aerospace Information Security and Trusted Computing, Ministry of Education, School of Cyber Science and Engineering, Wuhan University, Wuhan* 430079)

[2)](*Institute of Information Technology, Wuhan University, Rizhao, Shandong* 276800)

[3)](*Collaborative Innovation Center of Geospatial Technology, Wuhan* 430079)

Abstract In recent years, the widespread application of image classification models based on deep neural networks (DNNs) has significantly impacted critical fields, including facial recognition and autonomous driving. These models have showcased remarkable performance, revolutionizing the way we interact with technology. However, despite their success, deep neural networks are not without vulnerabilities, particularly in the face of adversarial attacks, which can lead to misclassification and compromise the integrity of these models. Addressing this challenge has become a

收稿日期:2023-06-29;在线发布日期:2023-12-29.本课题得到国家重点研发计划(2022YFB3102100)、国家自然科学基金(62206203,62076187)、湖北省重点研发计划(2022BAA039)、山东省重点研发计划(2022CXPT055)、武汉市科技计划(2023010302020707)资助.
李 瞧,博士研究生,主要研究方向为人工智能安全.E-mail:liqiaoqiao233@whu.edu.cn.**陈 晶**(通信作者),博士,教授,博士生导师,中国计算机学会(CCF)会员,主要研究领域为网络安全、分布式安全、区块链.E-mail:chenjing@whu.edu.cn.**张子君**,博士,副研究员,硕士生导师,主要研究方向为深度学习.**何 琨**,博士,副教授,硕士生导师,中国计算机学会(CCF)会员,主要研究领域为应用密码学、网络安全、云计算安全、人工智能安全、区块链安全.**杜瑞颖**,博士,教授,博士生导师,主要研究领域为网络安全、隐私保护.**汪欣欣**,博士研究生,主要研究方向为对抗攻击.

pivotal research direction, as ensuring the robustness of these models is essential for their real-world deployment. Currently, many defense methods, especially empirical ones, operate under the white-box assumption. This assumption relies on defenders having access to detailed information about the model, including its architecture and parameters. Unfortunately, model owners often hesitate to share such sensitive information due to privacy concerns. Even existing black-box defense methods struggle to provide comprehensive protection against attacks involving all norms, lacking the necessary universality. This inherent limitation has spurred the need for innovative solutions. In response to this challenge, this paper proposes a groundbreaking universal black-box certified defense method applicable to a broad spectrum of black-box models. The key innovation lies in the design of a query-based data-free substitute model generation scheme. Unlike traditional methods, this scheme eliminates the need for training data and prior knowledge of the model structure. Leveraging queries and zero-order optimization, it generates high-quality substitute models, effectively transforming the certified defense scenario into a white-box setting without compromising model privacy. Furthermore, this paper introduces additional layers of security through the incorporation of random smoothing and noise selection methods based on the white-box substitute model. These enhancements contribute to the construction of a universal certified defense solution capable of resisting adversarial attacks involving any norm. To validate the effectiveness of the substitute model, performance comparisons are made with the original model under white-box certified defense conditions. The experimental results, particularly on the CIFAR10 dataset, showcase the superiority of the proposed universal black-box certified defense solution over existing methods. The solution not only achieves significant improvements in certification accuracy but also maintains similar performance to white-box certified defense methods. Notably, compared to previous black-box certified defense methods, the proposed solution demonstrates over a 20% improvement in certification accuracy while effectively safeguarding the privacy of the original model. Specifically, the proposed solution successfully reduces the success rate of membership inference attacks by 5.48%, further highlighting its robustness and practical applicability in real-world scenarios.

Keywords　deep neural networks; certified defense; random smoothing; black-box models; substitute models

1　引　言

深度神经网络(Deep Neural Networks,DNNs)在人脸识别[1]、自动驾驶[2]等领域的应用已经取得了显著的效果.然而,DNNs 也面临着对抗样本攻击(Adversarial Examples,AEs)的挑战[3].对抗样本是通过对神经网络的输入数据进行微小的修改而产生的,这些修改在人眼中几乎不可察觉,但却足以使神经网络产生错误的输出结果.

对抗样本攻击在安全关键领域中可能带来严重后果.例如,在人脸识别系统中,恶意攻击者可以通过对人脸图像进行微小的修改来欺骗系统,导致身份识别错误[4].类似地,在自动驾驶系统中,对抗样本攻击可能误导车辆识别和决策,进而引发交通事故[5].

为了应对对抗样本的攻击,有研究人员从 DNNs 着手,旨在提升 DNNs 自身的鲁棒性.目前,已经提出了许多提升模型鲁棒性的方法.它们大致分为两类:经验防御和认证防御.经验防御指在现有攻击的基础上提出的防御策略,诸如对抗检测、对抗训练.然而,这些防御措施仍然可能受到二次攻击威胁,即攻击者专门设计来绕过防御措施的攻击.除此之外,该类方法缺少对模型鲁棒性的理论保证,即通过严格的数学分析和证明,确保模型对于输入数据的扰动和变化具有一定程度的稳定性和可靠性.为了能够证明模型在这些干扰下性能的下界或上界,得出模型的鲁棒性性质,有研究人员提出了认证防御.认证防御旨在为 DNNs 的鲁棒性提供一个可证

明的保证.它推导一个认证半径,对于任意的 L_p(如 L_1,L_2,L_∞)范数扰动,对测试输入添加的扰动不超过认证半径时,则 DNNs 的预测不能够被该扰动所干扰.即使存在二次攻击,只要添加的扰动不超过认证半径,依然可以保证预测的准确性.

虽然认证防御解决了鲁棒性的理论保证及二次攻击的问题.但是,现有的认证防御方法大多要求防御者在白盒模型上进行操作.模型所有者出于对模型细节的隐私和安全方面的考虑,他们可能不愿意分享这些细节.这种白盒假设会大大限制认证防御在实践中的应用.此外,有一个更为现实的问题是,有人需要大规模图像分类 API 提供一个鲁棒性更强的版本.这些问题皆表明,基于黑盒的认证防御在实践中的重要意义.

然而现有的黑盒认证防御方法只针对特定的设置,并不具有通用性,即不能对任何分类器上的任何输入进行鲁棒性认证,以应对由任何连续概率密度函数(Probability Density Function,PDF)产生噪声的对任意的 L_p范数扰动.

为了实现能够保护黑盒模型隐私的通用认证分类器,本文提出了一种新的策略,从一个固定的预训练分类器中获取一个认证分类器.鉴于黑盒下的认证防御不能够满足通用性的要求,本文将其转换为白盒场景,通过查询从预训练的分类器中蒸馏出一个替代模型,然后基于白盒的替代模型应用随机平滑构造通用认证分类器.在所有认证防御中,基于随机平滑的认证防御已经达到了最先进的认证半径,并且可以应用于任何分类器.随机平滑首先定义一个噪声分布,并将采样的噪声添加到测试输入中;然后在噪声输入的基础上建立一个平滑的分类器,最后得出认证半径.本文的策略利用了随机平滑的认证性质,以确保提出的方案是可证明的安全.

本文的贡献主要包括四个方面:

(1)本文设计了一种限制最小、先验知识最少的黑盒认证防御方法.本文将黑盒模型的认证防御问题定义为一种基于查询的白盒认证防御问题.将预训练的非鲁棒的 DNNs 转换为仅用查询功能就可认证的鲁棒模型.

(2)相较于以往基于黑盒的认证防御只能获取 L_2范数的认证半径,本文提出的通用黑盒认证方法可以基于任何分布的连续噪声来获得任意 L_p范数的认证半径.

(3)本文提出了基于替代模型的黑盒认证防御方法,通过生成替代模型,成功降低了成员推理攻击的成功率,提供了有效的隐私保护机制.实验结果表明,该方法在保持准确率的同时,有效降低了敏感信息的泄露风险.

(4)本文在多种模型和数据集上评估了提出的方案.实验结果显示,本文方案与白盒认证防御方法具有相似的效果,并相比之前的黑盒认证防御方法,在实现所有 L_p的认证防御的同时,准确率提升超过 20%.

2 相关工作

本节总结了提升深度神经网络鲁棒性的方法,主要分为经验防御和认证防御两类.这些方法旨在应对 DNNs 面临的对抗攻击,以提高其在恶意输入下的性能和安全性.

2.1 经验防御

经验防御是基于实践经验和已有攻击的基础上提出的防御策略.其中包括对抗检测和对抗训练.对抗检测[6-8]致力于识别输入中的对抗样本,以防止其进入 DNNs 并保持其预测的鲁棒性.对抗训练则通过在训练过程中引入对抗样本,让模型在处理具有噪声和扰动的输入数据时变得更加稳健和准确.Madry 等人[9]首次从理论上对对抗训练研究并且通过鲁棒优化的视角对其进行表述.对抗训练的出现使经验防御迅速发展.已有很多对抗训练方法从不同训练角度来提升模型的鲁棒性,诸如输入优化[10-11]或模型训练策略的优化[12-14].也有学者侧重于对抗训练原理的分析,探寻对抗训练提升模型鲁棒性的机理[15-16].

虽然上述方法在实践中已经取得了一定的成果,能够有效提升 DNNs 的鲁棒性,但是这些方法都是基于已有的攻击进行设计,因此很容易被专门设计来绕过防御的攻击的再次攻破.

2.2 认证防御

与经验防御不同,认证防御的目的是通过严格的数学分析和证明来提供对 DNNs 鲁棒性的理论保证.认证防御的目标是在给定的攻击模型和约束条件下,推导出 DNNs 对于输入数据的最坏情况下的性能下界或上界.这种方法能够为 DNNs 提供更可靠的鲁棒性保证,使其在面对各种攻击时能够保持高效和准确.一系列的认证防御方法被提出.

Lecuyer 等人[17]第一次提出了认证防御,利用差分隐私给分类器提供鲁棒证明.Cohen 等人[18]开发了基于随机平滑的认证防御,利用高斯噪声首次在 L_2范数的下提供了严格的鲁棒性保证.从此之

后，又有其他针对 L_p 范数的后续工作[19-21]，但这些方法都仅限于保证特定的 L_p 范数扰动的鲁棒性. Chen 等人[22]对随机平滑进行改进，保证了模型在 L_1，L_2，L_∞ 范数内的鲁棒性，但是一定程度提升了鲁棒认证的效率. 上述方法有效地提升了模型的鲁棒性，为模型提供了严格的可证明安全. 但是上述方法皆是基于白盒假设，十分依赖模型的先验知识(如模型架构、参数、数据集等). 这会大大限制认证防御在实践中的应用.

除此之外，近年来也产生了一些基于黑盒的认证防御工作. Salman 等人[23]率先进行了黑盒的认证防御，利用代理模型作为黑盒模型的近似值，在白盒设置之后可以对其进行防御. 然而，该方法不是严格意义上的黑盒设置，需要目标模型类型及其功能的信息等先验知识，且仅进行了 L_2 范数的验证. Zhang 等人[24]仅对目标分类器进行查询来构造认证分类器，但其仅能基于 L_2 范数的扰动进行防御，不具有通用性.

3　预备知识

3.1　随机平滑

随机平滑(Random Smoothing)在输入样本的周围添加随机噪声或扰动，使得输入样本的微小变化不会改变 DNNs 的预测结果. 通过在输入空间中对样本进行随机采样，并对采样结果进行平均，以获得平滑后的预测结果. 这种平均化的过程可以减少输入样本中的噪声和扰动对 DNNs 预测的影响，从而提高 DNNs 的鲁棒性.

随机平滑的基本思想是，为基础分类器生成一个平滑的分类器. 给定一个基础分类器 $f(x):\mathbb{R}^d\to\mathcal{Y}$，其中 $\mathbb{R}^d$ 是输入空间，$\mathcal{Y}$ 是输出空间. 那么平滑分类器 g 的定义如下：

$$g(x)=\arg\max_{c\in\mathcal{Y}}\mathbb{P}(f(x+\epsilon)),\ \epsilon\sim\mathcal{N}(0,\sigma^2 I)\quad(1)$$

对于输入 x，平滑分类器 g 会输出基础分类器 f 最有可能输出的类别概率. 基础分类器 f 对于输入 x 以 p_A 的概率输出最可能的类别 c，输出次一类别的概率为 p_B. 在分类任务下，对于 $\epsilon\sim\mathcal{N}(0,\sigma^2 I)$，假定 $c_A\in\mathcal{Y}$，且有 p_A、p_B，则有下式：

$$\mathbb{P}(f(x+\epsilon)=c_A)\geq p_A\geq p_B\geq\max_{c\neq c_A}\mathbb{P}(f(x+\epsilon)=c)\quad(2)$$

使用概率下界 $\underline{p_A}$ 替换 p_A，概率上界 $\overline{p_B}$ 替换 p_B，上式依然成立.

3.2　鲁棒性保证

认证半径是满足以下鲁棒性边界条件的所有对抗扰动 δ 的最小 L_p 范数. μ_x 是噪声 ϵ 的连续概率密度函数，t_A 和 t_B 是辅助参数，用于满足下述条件：

$$\begin{gathered}\mathbb{P}\left(\frac{\mu_x(x-\delta)}{\mu_x(x)}\leq t_A\right)=\underline{p_A},\\ \mathbb{P}\left(\frac{\mu_x(x-\delta)}{\mu_x(x)}\geq t_B\right)=\overline{p_B},\\ \mathbb{P}\left(\frac{\mu_x(x)}{\mu_x(x+\delta)}\leq t_A\right)=\mathbb{P}\left(\frac{\mu_x(x)}{\mu_x(x+\delta)}\geq t_B\right),\\ t_A=\Phi^{-1}(\underline{p_A}),\ t_B=\Phi^{-1}(\overline{p_B})\end{gathered}\quad(3)$$

其中，Φ^{-1} 是 $\mu_x(x-\delta)$ 的逆函数. 则平滑分类器 g 在 L_p 范数上的认证半径为

$$R=\frac{\sigma}{2}\cdot(t_A-t_B)\quad(4)$$

进而，对于任意的 $\|\delta\|_p<R$，则有 $g(x+\delta)=c_A$. 即对于任何可能的基础分类器 f 和任何可能的 x，平滑分类器 g 的输出值不会超过 $P(c_A=g(x))\pm R$.

由此，当噪声水平 σ 较高或最大类 c_A 的概率值较大，或其他类别的概率值较低时，鲁棒性半径 R 会变大. 当鲁棒性半径 $R\to\infty$ 时，则有 $p_A\to 1$ 且 $p_B\to 0$. 由于高斯分布存在于整输入空间 $\mathbb{R}^d$ 中，因此存在概率为 1 的情况，使得 $f(x+\epsilon)=c_A$.

3.3　问题定义

令 $f_b(x)$ 为一个预训练的黑盒预测模型，它可以将输入 x 映射到一个预测结果. 本文所研究的黑盒场景是在 f_b 的所有者无法共享模型细节和训练数据的情况下进行黑盒防御. 因此，与黑盒模型唯一的交互方式是提交一个输入并接收相应的预测输出. 本文问题的形式化定义如下：对于给定的黑盒基础模型 f_b，仅通过输入输出函数 $Q(f_b)$ 查询来设计一个认证分类器 D，以产生对抗攻击具有鲁棒性的模型 $D(x)$.

4　基于替代生成的无数据黑盒认证防御方案

基于替代生成的无数据黑盒认证防御方案的核心思想是将黑盒场景转换为白盒场景，然后在白盒场景下，进行基于随机平滑的认证防御. 本节首先介绍了将黑盒问题转化为白盒问题的核心方法-基于查询的无数据替代模型生成方案. 然后介绍基于替代模型的黑盒认证防御方案的详细设计过程和细节.

4.1　基于查询的无数据替代模型生成方案

无数据替代模型生成的目标是训练一个替代模型 S，用于拟合黑盒模型 B 在其输入空间 D_b 上的预测. 换言之，就是要找到替代模型参数 θ_S，对于输入 x，使得替代模型 $S(x)$ 与 $B(x)$ 在输入空间 D_b 上的

误差最小化.则对于所有的 $x\in D_b$:

$$\arg\min_{\theta_s}\mathbb{P}_{x\sim D_b}[\arg\max B(x)\neq\arg\max S(x)] \quad (5)$$

由于黑盒模型的输入空间 D_b 是不公开的,所以本文使用的无数据替代模型生成方案通过一个生成的数据集 D_s 最小化替代模型的误差来实现.通过优化损失函数来评估替代模型和黑盒模型的不一致性:

$$\arg\min_{\theta_s}\mathbb{E}_{x\sim D_s}[L(S(x),B(x)] \quad (6)$$

本小节将介绍方案如何基于最小化的查询预算生成替代模型的训练集以及如何训练替代模型.

4.1.1 模型设计

无数据的替代模型生成方案与 GAN 类似.生成器 G 负责生成替代模型的训练数据,而替代模型 S 用来充当鉴别器,来匹配目标模型 B 对生成器生成的数据的预测.在这种情况下,两个对手是 S 和 G,他们分别试图最小化和最大化 S 和 B 之间的分歧.

替代模型生成方案如图 1 所示.首先从标准正态分布中采样随机噪声 z,并将其输入生成器 G,生成对应的图像.然后,黑盒模型 B 和替代模型 S 分别对这个生成图像进行推理.最后,计算相关的损失函数 Loss.在计算来与替代模型参数 θ_S 相关的梯度和与生成器参数 θ_G 相关的梯度,进行反向传播.由于目标模型是黑盒访问的模式,不能够进行梯度的传播,需要进行梯度近似.

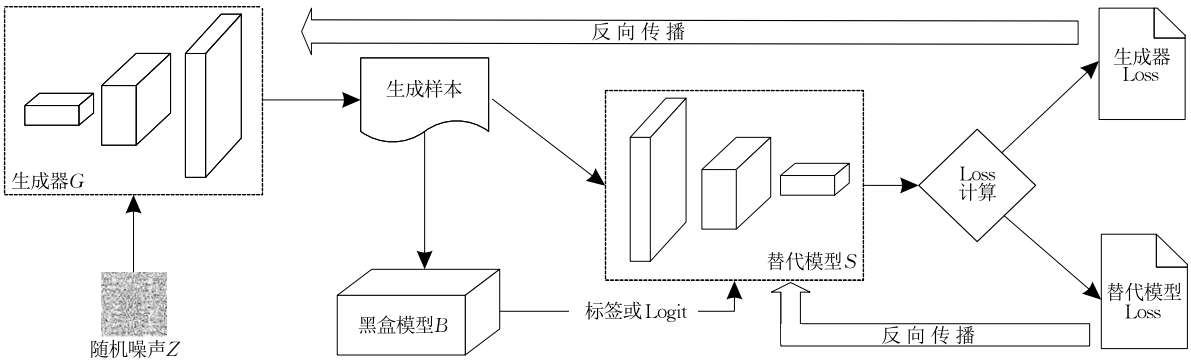


图 1 替代模型生成方案

生成器 G 的目的是生成图像,使 S 和 B 之间的差异最大化.生成器 G 的损失函数与替代模型 S 相同.不同的是,替代模型 S 的目标是最小化损失函数,而 G 想要最大化它.换句话说,替代模型 S 的目的是要匹配黑盒模型的预测,而生成器 G 的训练目标是生成对于替代模型困难的样本.这样就形成了一个类似于 GAN 的博弈对抗过程,该过程的描述如下式所示:

$$\arg\min_{S}\arg\max_{G}\mathbb{E}_{z\sim\mathcal{N}(0,1)}[L(B(G(z)),S(G(z)))] \quad (7)$$

4.1.2 模型训练

每一轮迭代中需要交替训练生成器 G 和替代模型 S.为了使生成 G 和替代模型 S 训练之间的平衡方便调整,在进入下一个迭代之前,分别重复训练 n_G 和 n_S 次.设置一个较高的 n_G 能够使 G 的训练速度加快,并且生成对于 S 更困难的样本.

损失函数. 损失函数对替代模型生成的效果至关重要.错误的损失函数会导致梯度消失,影响优化器的收敛.为了防止梯度消失问题的出现,本文使用范数损失作为损失函数.范数损失是一种度量两个向量之间差异的方法,它是指向量中各个元素绝对值之和,可以用于衡量替代模型和目标模型输出之间的差异.尽管范数在所有地方都不可微分,但在实践中它不会遭受梯度消失的问题,并且在实验中取得更好的结果.在本文中,使用损失函数如下:

$$L_1(x)=\sum_{i=1}^{M}|b_i-s_i|,\ i\in 1,\cdots,M \quad (8)$$

梯度估计. 由于黑盒模型 B 仅能够进行查询,为了训练生成器 G,必须计算相对于 G 的参数 θ_G 的梯度 $\nabla_{\theta_G}L$.因此,需要通过与黑盒模型进行交互来近似梯度:本文使用零阶优化来近似黑盒模型的梯度,用以训练生成器 G.其近似过程如下式所示:

$$\hat{\nabla}B(x)=\frac{1}{N}\sum_{i=1}^{N}\frac{B(x_i)-B(x)}{h} \quad (9)$$

4.2 基于替代模型的通用黑盒认证防御方案

通用的认证防御需要能够在由任意噪声概率密度分布构造的鲁棒边界内,普遍地、自动地推导出针对任意 L_p 扰动的鲁棒半径.本小节将介绍方案如何基于替代模型获得通用的黑盒认证防御.

4.2.1 基于随机平滑的认证防御

基于随机平滑认证分类器 $G(x)$ 需要在类分布 $f(x+\epsilon)$ 具有最大概率的类 c_A.

训练过程. 基于随机平滑的认证防御的分类器在训练过程中使用噪声采样来近似样本的类别分

布，然后根据出现次数最多的类别作为预测结果. 如果其他类别的出现次数超过设定的阈值，则返回弃权结果. 预测过程通过基分类器运行 x 的 n 个噪声来采样 n 个 $f(x+\epsilon)$ 样本. 令 c_A. 为出现次数最多的类别. 如果 c_A 出现的频率比任何其他的类别都要多，则 PREDICT 返回 c_A. 认证分类器使用假设检验来校准弃权阈值，以便将返回错误类别的概率限制为 α，具体流程如算法 1 所示.

算法 1.　认证分类器训练.

输入：训练集 X，标签集 C，基础分类器 f，蒙特卡洛采样数 n，弃权阈值 ζ

输出：认证分类器 D

TRAIN

1. FOR x_i, c_i IN X, C:
2. 　根据样本的标签 c_i 计算类别分布 $f(x_i+\epsilon)$
3. 　噪声采样，通过采样 n 个噪声样本来近似类分布 $f(x_i+\epsilon)$
4. 　统计样本中每个类别的出现次数 *counts*
5. 　取 *counts* 中出现次数最多的类别 c_A
6. 　FOR 每个类别 c:
7. 　　IF $c \neq c_A$ 且 $counts[c] > counts[c_A] * \zeta$
8. 　　更新 D
9. RETURN D

PREDICT

10. 噪声采样，采样 n 个噪声样本来近似类分布 $f(x+\epsilon)$
11. 从类分布 $f(x_i+\epsilon)$ 中采样一个样本 x_i
12. 取 *counts* 中出现次数最多的类别 c_A
13. FOR 每个类别 c:
14. 　IF $c \neq c_A$ 且 $counts[c] > counts[c_A] * \zeta$
15. 　　RETURN 弃权
16. RETURN c_A

认证过程. 在模型的预测过程中，会对某些样本给出不确定性分类. 通过认证过程，可以检测出模型对于样本是否具有高置信度. 若样本的置信度不足，则模型不对该样本进行分类，即为对该样本弃权(*abstain*). 该过程通过比较出现次数最多的类别的频率与认证阈值来确定是否认证通过. 具体流程如算法 2 所示.

算法 2.　输入认证.

输入：测试样本 x，基础分类器 f，蒙特卡洛采样数 n，弃权阈值 ζ，认证阈值 ι

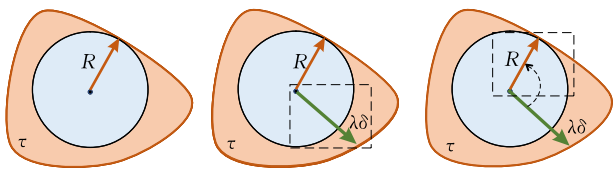
输出：认输入 x 的认证结果，c_A 或弃权

1. 噪声采样，采样 n 个噪声样本来近似类分布 $f(x_i+\epsilon)$
2. 统计样本中每个类别的出现次数 *counts*
3. 取 *counts* 中出现次数最多的类别 c_A
4. FOR 每个类别 c:
5. 　IF $c \neq c_A$ 且 $counts[c] > counts[c_A] * \zeta$
6. 　　RETURN 弃权
7. 计算类别 c_A 的频率 $counts[c_A]/n$
8. IF c_A 的频率 $> \iota$
9. 　RETURN c_A
10. ELSE
11. 　RETURN 弃权

基于随机平滑的认证防御可以提高分类器对抗样本的鲁棒性，减少错误分类的可能性，并提供一定程度的认证保证. 其认证半径的推导及噪声采样过程在下面介绍.

4.2.2　鲁棒半径推导

如图 2 所示，τ 是一个稳健的空间，只要添加的扰动仍在 τ 空间内，平滑分类器的预测就可以被证明是正确的. 对于一个输入样本，其紧密的认证半径 R 可以通过在鲁棒边界上找到一个 $\|\delta\|_p$ 最小的扰动 δ 推导得出. 然而，找到一个恰好在鲁棒性边界上的扰动 δ，或者找到最小值 $\|\delta\|_p$ 是困难的. 因此，本文使用 Hong 等人[25]提出的两阶优化方案来获得紧密的认证半径. 第一阶段的目的是使得 δ 位于鲁棒性边界上，第二阶段的目的是最小化 L_p 范数.



(a) 稳健空间τ　(b) 第一阶段：标量优化　(c) 第二阶段：方向优化

图 2　鲁棒半径推导示意图

如图 2 所示，在第一阶段进行标量优化，找到一个缩放因子 λ，将扰动 δ 缩放到鲁棒边界上；在第二阶段进行方向优化，δ 的方向将被优化以达到最小化 $\|\lambda\delta\|_p$. 在这两阶段优化中，方向优化将被迭代执行，直到找到最小的 $\|\lambda\delta\|_p$，其中在每次方向迭代过程中，扰动 δ 将被事先缩放到鲁棒性边界上. 因此，对于第 3.2 节中式(3)中的难以处理的优化问题可以转化为

$$
\begin{aligned}
&R = \|\lambda\delta_p\|,\\
&\text{s.t. } \delta \in \arg\min_{\delta} \|\lambda\delta\|_p,\ \lambda = \arg\min_{\lambda} |K|,\\
&\mathbb{P}\left(\frac{\mu_x(x-\lambda\delta)}{\mu_x(x)} \leq t_A\right) = \underline{p_A},\\
&\mathbb{P}\left(\frac{\mu_x(x-\lambda\delta)}{\mu_x(x)} \geq t_B\right) = \overline{p_B},\\
&K = \mathbb{P}\left(\frac{\mu_x(x)}{\mu_x(x+\lambda\delta)} \leq t_A\right) - \mathbb{P}\left(\frac{\mu_x(x)}{\mu_x(x+\lambda\delta)} \geq t_B\right),\\
&t_A = \Phi^{-1}(\underline{p_A}),\ t_B = \Phi^{-1}(\overline{p_B})
\end{aligned}
\tag{10}
$$

式(10)旨在找到一个缩放因子 λ,将扰动 δ 缩放到边界,使得 $|K|$ 接近于 0.通过使用标量 λ 确保缩放后的 δ 几乎在边界上;方向优化扰动 δ 的方向以找到认证半径 $R=\|\lambda\delta\|_p$.

根据现有的基于随机平滑的防御,首先使用蒙特卡罗方法[18]估计概率上界 $\underline{p_A}$ 和 $\overline{p_B}$;然后,将其应用到两阶优化方案中.来推导认证半径.

估计概率边界. 两阶段优化需要估计概率上下界 $\underline{p_A}$ 和 $\overline{p_B}$,并计算两个辅助参数 t_A 和 t_B.

概率上下界 $\underline{p_A}$ 和 $\overline{p_B}$ 是通过蒙特卡洛方法[18]进行估计的.给定估计的 $\underline{p_A}$ 和 $\overline{p_B}$,以及任何给定的噪声概率密度函数和扰动 δ,使用蒙特卡洛方法来估计分数 $\frac{\mu_x(x-\lambda\delta)}{\mu_x(x)}$ 的累积密度函数.接下来计算辅助参数 t_A 和 t_B.具体而言,辅助参数 t_A 和 t_B 可以通过 $t_A=\Phi^{-1}(\underline{p_A})$ 和 $t_B=\Phi^{-1}(\overline{p_B})$ 来计算,其中 Φ^{-1} 是分数 $\frac{\mu_x(x-\lambda\delta)}{\mu_x(x)}$ 的反函数累积密度函数.计算 t_A 和 t_B 的步骤详见算法 3.

算法 3. 输入认证.

输入:概率的下界 $\underline{p_A}$,概率的上界 $\overline{p_B}$,扰动的标量值 λ,扰动 δ,噪声的概率密度函数 μ_x,蒙特卡洛方法中的样本数 n

输出:辅助参数 t_A,t_b

1. 从离散版本的概率密度函数中采样 n 个噪声 $=\epsilon\in\mathbb{R}^{n\times d}$.
2. 使用这些 n 个噪声样本,μ_x,λ 和 δ 计算 $\mu_x(x-\lambda\delta)$.
3. 使用蒙特卡洛方法估计 $\mu_x(x-\lambda\delta)$ 的累积分布函数,记为 Φ.
4. RETURN $tA=\Phi^{-1}(\underline{p_A})$ 和 $tB=\Phi^{-1}(\overline{p_B})$,其中 Φ^{-1} 是 $\mu_x(x-\lambda\delta)$ 的逆函数.

标量优化. 令 $|K|$ 为扰动 δ 与鲁棒性边界之间的距离.本文使用二分搜索法来寻找最小化 $|K|$ 的缩放因子.当 $K=0$ 时,扰动 δ 正好位于鲁棒性边界上.在固定 δ 的方向的情况下,需要寻找两个标量使得 $K>0$ 和 $K<0$.首先基于标量 λ_a 来计算 K.如果 $K>0$,则缩放后的扰动 $\lambda_a\delta$ 位于鲁棒性边界内.由此,可以增大标量以找到 λ_b 使得 $K<0$,反之亦然.然后,使用 $\lambda=\frac{1}{2}(\lambda_a+\lambda_b)$ 迭代计算 K.如果 $K>0$,则令 $\lambda_a=\lambda$;否则,令 $\lambda_b=\lambda$.重复这个迭代过程,直到 K 小于一个阈值或者迭代次数足够大.具体步骤如算法 4 所示.

算法 4. 输入认证.

输入:概率的下界 $\geq p_A$,概率的上界 $\geq p_B$,扰动的标量值 λ,扰动 δ,噪声概率密度函数:μ_x,蒙特卡洛方法中的样本数量 n,K 的阈值:K_m,二分搜索的迭代次数:N

输出:使 $|K|$ 最小的标量 λ

1. 找到初始标量 λ_a 和 λ_b,使得 $K>0$ 和 $K<0$
2. $\lambda=(\lambda_a+\lambda_b)/2$
3. 使用 λ 计算 K
4. WHILE $N>0$ 且 $|K|>K_m$ DO
5. IF $K>0$ THEN
6. $\lambda_a=\lambda$
7. ELSE
8. $\lambda_b=\lambda$
9. $\lambda=(\lambda_a+\lambda_b)/2$
10. 使用 λ 计算 K
11. $N=N-1$
12. RETURN λ

方向优化. 本文使用粒子群优化(Particle Swarm Optimization,PSO)方法来寻找在缩放到鲁棒性边界后最小化 L_p 范数的 θ.在 PSO 的每一次迭代中,粒子的位置表示 δ,成本函数为 $f_{\text{PSO}}(\delta)=\|\lambda\delta\|_p$,其中 λ 通过标量优化得到.PSO 旨在找到能够最小化成本函数的 δ.

4.2.3 噪声采样

方案设计. 现有的随机平滑方法是使用相同的噪声(一般为高斯噪声)来训练平滑分类器来证明分类器的鲁棒性,这是由于这样的训练可以提升预测概率的下界.本文将噪声的概率密度分布视为变量,来最大化认证半径.不同噪声概率密度分布产生的认证半径不同,分类器为了更好地防御 L_p 扰动,可以将噪声概率密度分布视为一个变量,寻找使分类器认证半径最大的最优噪声概率密度分布,并从该噪声概率密度分布中注入相同的噪声到训练数据中以训练分类器和测试输入中以构建平滑分类器.

本文提出的方案可以自动逼近认证半径,根据不同的噪声概率密度分布评估认证防御的整体性能.将噪声概率密度分布表示为 μ,最优噪声的选择问题可以定义为找到一个可以定义为分类器找到一个 μ 使得认证分类器的整体性能最优.在进一步调整每个输入或分类器的噪声概率密度分布过程,具体地说,设 $\mu(x,\alpha)$ 表示噪声概率密度分布,其中 α 是函数中的一组超参数,即 $\alpha=[\alpha_1,\alpha_2,\cdots,\alpha_m]$.最优噪声的选择使用网格搜索来搜索噪声概率密度分布的最佳参数.在算法执行过程中,如果在每轮中找到更

好的解决方案，则迭代更新输入的超参数，直到收敛.

通用性说明. 拉普拉斯噪声和高斯噪声是一般正态分布$\propto e^{-|x/\alpha|^{\beta}}$的特殊形式. 没有理论可以表明高斯噪声和拉普拉斯噪声对于L_1和L_2扰动是最优的. 有文献指出有一些更好的噪声可以应用于L_1和L_2的扰动. 本文方案将找到生成最佳噪声的最优参数α和β，以最大化每个L_p扰动认证半径.

5　实验结果与分析

本节主要以具体实验论证 NBCD 方法的有效性和优越性. 首先，介绍用于评估 NBCD 方法的实验环境和参数设置；然后，分别对替代模型的效果进行分析与实验；接下来，对 NBCD 的训练和测试结果进行分析，并与先进的基线实验进行比较验证. 之后，对模型的隐私性进行的测试，验证 NBCD 对于黑盒模型的隐私性保护；最后进行消融实验分析，验证 NBCD 的有效性.

5.1　数据集与评估指标

数据集. 实验使用 CIFAR10[26] 和 SVHN[27] 上评估方案的性能，这些数据集是评估图像分类和认证防御的常用数据集.

评估指标. 对于替代模型的性能评估，实验使用准确率和比率两个指标. 替代模型准确率是指对于基础模型的测试集在替代模型中的准确率. 比率指将原始模型的准确率视为 1，替代模型准确率与原始模型准确率之间比值，该指标可以有效反映出替代模型的替代效果.

在认证防御中，常使用认证准确率进行评估. 认证准确率是认证分类器的可证明鲁棒保证. 实验使用[18]中提供的方法计算与输入样本相关的L_2范数扰动的认证半径. 给定半径R，认证准确率是正确分类的数据中认证半径大于R的百分比.

如图 3 所示，文献[28]绘制了在不同方差的噪声的认证准确率与认证半径的相关曲线，该曲线表示认证半径范围内认证准确率.

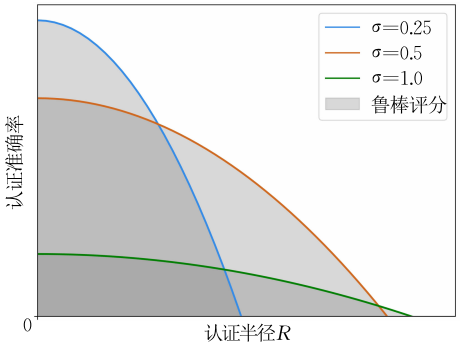


图 3　认证防御整体性能

为了衡量整体性能，实验使用曲线下的面积作为认证鲁棒性的总体度量，该度量指标的定义如下：

$$\int_0^{+\infty} \max_{\sigma}(Acc_{\sigma}(R))\mathrm{d}R,\ \sigma\in\Sigma \tag{11}$$

其中$Acc_{\sigma}(R)$是由方差为σ的噪声计算的半径R处的认证准确率，Σ是一组候选σ.

同时，本文也绘制了P_A-R曲线图评估认证半径的质量. P_A-R曲线图是认证防御中常用的评估指标，用于衡量模型对于输入样本的分类置信度和认证半径之间的关系. P_A代表对于一个输入样本来说最可能被分类到的类别c_A的置信度，而R表示模型在输入样本周围容忍的错误分类距离范围，即认证半径.

5.2　替代模型训练与测试

知识蒸馏中有研究表明[29]，一个较小的学生模型足以学习到一个较大的替代模型的知识. 因此，实验使用 ResNet18 作为替代模型架构.

替代模型的训练采用的 batch 大小为 256. 在实验中，SVHN 数据集的默认查询预算为 2M，CIFAR10 数据集为 20M. 生成器使用了三个卷积层，其中插入了线性上采样层、批归一化层和 ReLU 激活函数，除最后一层外，所有层都使用 ReLU 激活函数. 最后一层的激活函数是双曲正切函数，用于将输出值限定在范围[−1，1]内. 对于梯度近似，实验采样$m=1$个随机方向，并使用步长$\varepsilon=10^{-3}$.

首先，实验比较了在默认查询预算下的替代模型生成方案的效果. 如表 1 和表 2 所示，实验比较了对于不同数据集在不同原始模型架构上的性能. 实验评估了替代模型与原始模型测试集上的准确率与原始模型准确率. 除此之外，实验使用比率这一指标描述替代模型与原始模型之间的关系，替代模型与原始模型准确率之间的比值即为比率.

表 1　CIFRA10 中无数据替代模型生成方法准确率

原始模型架构	原始模型准确率/%	替代模型准确率/%	比率/%
Resnet18	93.01	87.17	93.73
Resnet34	93.15	88.31	94.08
Resnet50	95.63	88.45	94.05

表 2　SVHN 中无数据替代模型生成方法准确率

原始模型架构	原始模型准确率/%	替代模型准确率/%	比率/%
Resnet18	90.32	89.45	99.03
Resnet34	87.51	87.89	100.40
Resnet50	86.21	86.04	99.80

此外，实验还使用了不同的查询预算生成替代模型，生成模型的准确率如图 4 所示.

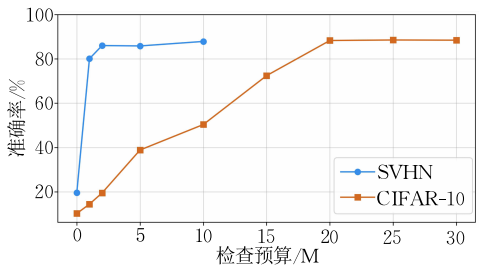


图 4 不同查询预算下替代模型准确率

实验表明，在没有任何关于原始模型的先验知识情况下，在原始模型架构为 Resnet34 时，该替代模型生成方法在查询预算为 20 M 时达到了 88.31% 的准确率，约 0.94 的比率，并且在查询预算为 25 M 时达到了 88.55% 的准确率，约 0.95 的比率. 对于 SVHN 的原始模型，可以观察到类似的结果：仅使用 2 M 的查询预算就能达到 86.04% 的准确率，约 0.98 的比率. 并且在查询预算为 10 M 时达到了 1 的比率. 同样的，在其他模型架构中也有着类似的结果. 实验结果表明，本文提出的替代模型生成方案能够在不同的模型架构中生成与原始模型相似的替代模型.

5.3 基于替代模型的通用黑盒认证防御评估

在本节中，实验将全面评估 NBCD 方案，并且使用最先进的黑盒认证防御进行对比. 首先，实验将使用不同的概率密度函数的噪声，在不同 L_1，L_2，L_∞ 扰动下的近似认证半径评估 NBCD 的通用性. 其次，实验将与现有的工作的认证半径相比较，评估本文的先进性. 最后，实验进行了基线实验，比较了先进的黑盒认证防御上的认证准确率.

5.3.1 参数设置

实验使用网格搜索的方式来搜索噪声概率密度分布的最佳参数. 对于服从 $\propto e^{-|x/\alpha|^\beta}$ 的噪声，首先选择作为主要参数，并设置 α 以满足 $\sigma=1$.

实验在 SVHN 数据集上，每一轮网格搜索都训练一个模型，然后对测试集中的一组图像进行认证. 对每一对参数 α 和 β，基于随机平滑训练一个多层感知器. 最后在一组 $\sigma=[0.25,0.50,0.75,1.00]$ 上计算鲁棒评分，进行整体性能评估. 其中，在计算认证半径时，蒙特卡洛采样数设置为 1000，其结果如图 5 所示.

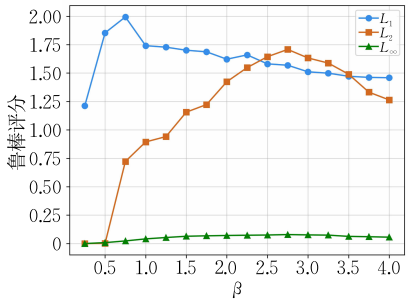


图 5 不同噪声分布下的鲁棒评分

从图 5 可以看出，对于 L_1 范数，其最佳的 β 值为 0.75；对于 L_2 与 L_∞ 范数来说，最佳性能的 β 为 2.25. 由此可以得出，对于 L_1，L_2 和 L_∞ 范数来说，常用的，如 $\beta=1$ 的拉普拉斯噪声与 $\beta=2$ 高斯噪声并不是认证防御中最优的噪声. 较小的 β 可以在认证半径和认证准确率之间提供更好的平衡.

5.3.2 通用性评估

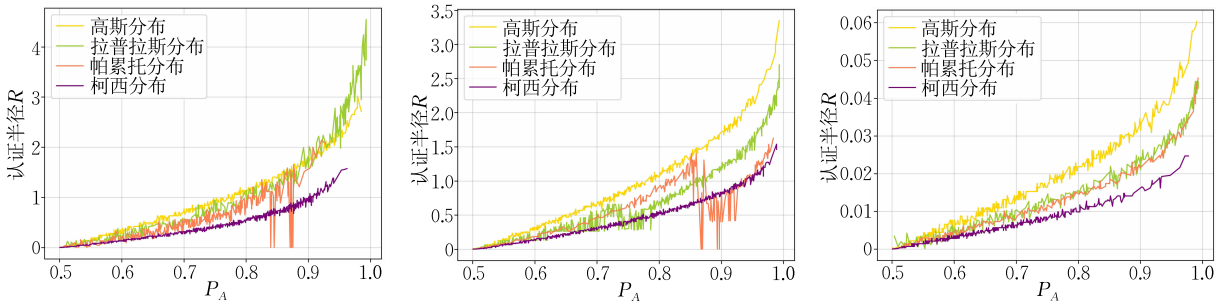
由于随机平滑对任何输入和任何分类器都具有经过认证的鲁棒性，由此，实验评估的目标是，对于任意的概率密度函数的噪声和任意的 L_p 扰动.

为了找到对每个 L_p 扰动最优的概率密度函数，实验计算了不同概率密度函数分布，具有相同方差的认证半径. 在 $P_A\in(0.5,1.0]$ 时计算认证半径. 具体噪声分布如表 3 所示.

表 3 噪声分布

噪声分布类型	概率密度函数
高斯分布	$\propto e^{-\lvert x/\alpha\rvert^2}$
拉普拉斯分布	$\propto e^{-\lvert x/\alpha\rvert}$
帕累托分布	$\propto \frac{1}{(1+\lvert x/\alpha\rvert)^{\beta+1}}$
柯西分布	$\propto \frac{\alpha^2}{x^2+\alpha^2}$

如图 6 所示，本文绘制了表 3 所列出的噪声分布



(a) L_1范数 (b) L_2范数 (c) L_∞范数

图 6 不同噪声分布的 P_A-R 曲线图

对于L_1,L_2和L_∞的P_A-R 曲线图. 对于所有的L_p扰动,高斯噪声在大多数P_A上生成的认证半径都是最大的. 除了柯西分布外,所有的曲线非常接近. 当P_A对于L_2和L_∞扰动较低时,方案无法找到基于拉普拉斯分布的认证半径. 有研究表明,使用Laplace噪声用于认证防御难以推导出针对L_2和L_∞扰动的认证半径.

5.3.3　对比实验

在本节中,实验在CIFAR10数据集上,比较了本文方案与现有的黑盒认证防御方案. 对比实验方案使用Zhang等人[24]提出的方案作为基线方法. Zhang等人提出的方法是严格的仅使用查询的黑盒场景,但其仅能在L_2范数上进行认证防御. 因此仅在L_2范数上进行该方案的对比实验. 除此之外,实验将原始的黑盒模型视为白盒模型,使用Cohen等人[18]提出的方案进行白盒模型上的认证防御,作为一个基线实验. 由于本文提出的方案是没有原始模型训练集的先验知识,因此使用第4.1节中所训练得到的生成器G生成训练数据.

在CIFAR10数据集上,实验使用了结构为ResNet34的分类器作为黑盒分类器,如表4所示,原始模型的准确率为93.15%. 本文方案和Zhang等人的方法中,实验使用同样的查询预算,均设为20 M.

表4　不同认证防御方法在CIFAR10数据集中认证准确率及鲁棒评分对比

L_1半径	0.25	0.50	1.00	1.50	2.00	2.50	鲁棒评分
Cohen's	72.95	60.47	48.86	36.75	30.77	26.75	0.9967
Base	77.37	65.06	48.93	40.02	39.90	30.92	1.2122
NBCD	76.98	64.65	49.03	40.02	34.26	29.94	1.0300
L_2半径	0.25	0.50	1.00	1.50	2.00	2.50	鲁棒评分
Zhang's	12.61	12.65	11.68	11.16	10.23	9.98	0.2535
Cohen's	75.41	65.20	45.40	38.20	33.50	28.90	0.9974
Base	77.52	64.94	49.39	40.16	34.27	29.97	1.0345
NBCD	76.52	54.75	49.51	39.97	33.97	30.67	1.0324
L_∞半径	0.25	0.50	1.00	1.50	2.00	2.50	鲁棒评分
Cohen's	77.02	64.38	48.29	38.75	32.27	29.84	1.0088
Base	77.03	65.09	48.89	39.75	34.10	29.77	1.0285
NBCD	76.20	64.64	49.17	39.60	34.05	30.07	1.0269

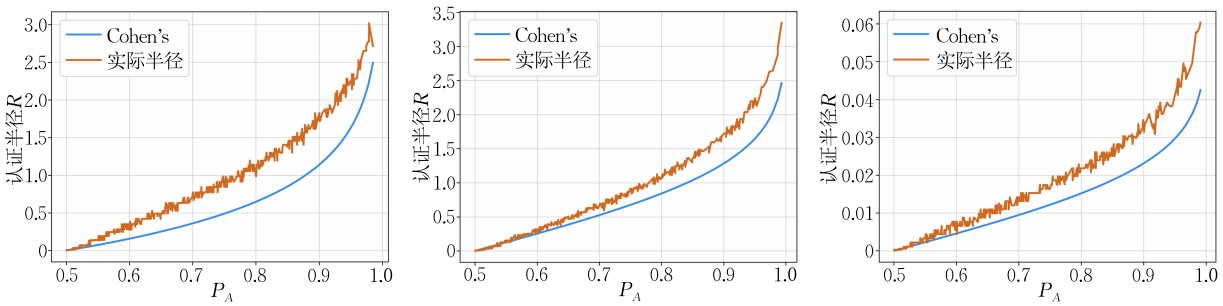
对于CIFAR10,实验使用一系列方差为σ的噪声PDF来计算鲁棒评分,评估方案的整体性能.σ的取值范围为[0.25,0.5,0.75,1.0]. 噪声分布的初始化设置与基准方法相同,拉普拉斯噪声的β值为1,高斯噪声的β值为2.

实验使用拉普拉斯分布来抵抗L_1扰动,使用高斯噪声测试L_2和L_∞扰动,对于本文和基线方法,蒙特卡洛采样数均设为4000. 认证准确率是通过对测试集中随机选择的500个图像的认证半径进行计算的.

根据表4的实验结果,在CIFAR10数据集上,可以明显观察到NBCD方案相对于基于黑盒模型的认证方法[24],在认证准确率和鲁棒评分方面取得了显著的提高. 与Zhang等人提出的方案相比,NBCD在L_2范数上的鲁棒评分提高了77.89%.

与Cohen的白盒方法相比,尽管NBCD方案的认证准确率略有下降,但其鲁棒评分仍然相近甚至有所提高,在L_1,L_2和L_∞范数下的鲁棒评分分别提高了6.33%、3.50%和1.81%. 这表明NBCD方案能够更有效地抵抗对抗攻击,提升模型的鲁棒性.

本文还与Cohen的白盒方法在所有的L_p范数上进行了鲁棒半径的比较,并绘制了P_A-R曲线图. 从图7可以观察到,本文方案所获得的认证半径优于Cohen方案的结果. 这些结果表明,本文方案能



(a) L_1范数　(b) L_2范数　(c) L_∞范数

图7　NBCD与先进方案的P_A-R曲线图的比较

够对具有任何连续噪声分布的输入数据，在任意 L_p 范数下实现更好的鲁棒认证性能.

因此，实验结果表明 NBCD 方案在认证防御中具有明显的优势，并显示出在实际应用中具备较强的可行性.

5.4 消融实验

为了深入研究替代模型对认证防御的影响，本文设计了一系列的消融实验，以进一步验证并探究所提出的 NBCD 方案的有效性.

本文设计了基准方案 Base 作为消融实验，直接将原始模型视为白盒系统应用第 4.2 节中提出的基于随机平滑的通用认证防御方案. 消融实验评估了 NBCD 方案与基于随机平滑的方案之间的差异，并了解替代模型的生成对认证防御的影响.

从表 4 中可以观察到，在所有的认证半径和范数下，NBCD 方案的认证准确率和鲁棒评分与基于随机平滑的方案非常相似. 这意味着通过采用替代模型进行认证防御，NBCD 方案能够达到与基准方案相当的效果，而替代模型的生成对防御结果几乎没有产生明显的影响.

5.5 隐私性测试

为了验证替代模型生成方案对原始模型隐私性的有效保护，本文利用 Song 等人[30]提出的方法进行了成员推理攻击实验来进行方案的隐私性测试. 成员推理攻击常用于评估模型的隐私性，通过尝试从替代模型中恢复原始模型的训练数据来测试其泄露敏感信息的能力. 结果如表 5 所示.

表 5 隐私性测试

模型	准确率/%	成员推理攻击成功率/%
原始模型	91.02	56.94
替代模型	87.16	53.98
Base	64.77	54.45
NBCD	64.16	51.48

如表 5 所示，原始模型的高准确率表明其在任务上取得了很好的性能，但成员推理攻击成功率高达 56.94%，暴露了隐私泄露的风险. 基于原模型的白盒认证防御方法在准确率上进一步下降，但成员推理攻击成功率保持在相似水平. 虽然，替代模型在准确率上有所下降，同时成员推理攻击成功率也有所降低，表明生成替代模型在一定程度上降低了攻击者从模型中恢复训练数据的能力，从而提供了一定程度的隐私保护.

基于生成替代模型的 NBCD 方法在保持了和白盒认证防御相似的准确率的同时，在成员推理攻击的防御方面都取得了显著的改进，使成员推理攻击的成功率降低了 2.97%. 较低的成员推理攻击成功率表明攻击者在恢复训练数据方面面临更大的困难. 由于替代模型与黑盒模型的模型结构及训练集不同，因此增加了成员推理攻击的攻击难度.

实验表明，利用替代模型的黑盒认证防御是一种有效的隐私保护策略. 该策略增加了攻击者的困难，提供了更可靠的隐私保护机制，能够有效防止敏感信息的泄露.

6 结束语

本文提出了一种基于黑盒模型的认证防御方法. 这一方法仅通过查询就能够有效的为黑盒模型提供通用的认证防御. 与其他黑盒认证防御方法不同，NBCD 通过无数据蒸馏出黑盒模型的替代模型，将其转换成白盒场景进行认证防御. 我们的实验表明，NBCD 蒸馏出的替代模型能有有效近似黑盒模型，基于此蒸馏模型设计的认证防御分类器能有有效的在任何噪声分布的情况下实现对任意范数的防御，实现了黑盒认证防御通用性的目标.

参 考 文 献

[1] Wang M, Deng W. Deep face recognition: A survey. Neurocomputing, 2021, 429: 215-244

[2] Grigorescu S, Trasnea B, Cocias T, et al. A survey of deep learning techniques for autonomous driving. Journal of Field Robotics, 2020, 37(3): 362-386

[3] Szegedy C, Zaremba W, Sutskever I, et al. Intriguing properties of neural networks. CoRR, abs:1312.6199, 2013

[4] Jia S, Yin B, Yao T, et al. Adv-attribute: Inconspicuous and transferable adversarial attack on face recognition. Advances in Neural Information Processing Systems, 2022, 35: 34136-34147

[5] Zhang Q, Hu S, Sun J, et al. On adversarial robustness of trajectory prediction for autonomous vehicles//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA, 2022: 15159-15168

[6] Xu W, Evans D, Qi Y. Feature squeezing: Detecting adversarial examples in deep neural networks//Proceedings of the 2018 Network and Distributed System Security Symposium. San Diego, USA, 2018

[7] Geifman Y, El-yaniv R. SelectiveNet: A deep neural network with an integrated reject option//Proceedings of the International Conference on Machine Learning. Long Beach, USA, 2019: 2151-2159

[8] Tian J, Zhou J, Li Y, et al. Detecting adversarial examples from sensitivity inconsistency of spatial-transform domain//Proceedings of the AAAI Conference on Artificial Intelligence. Virtual, 2021: 9877-9885

[9] Madry A, Makelov A, Schmidt L, et al. Towards deep learning models resistant to adversarial attacks//Proceedings of the International Conference on Learning Representations. Vancouver, Canada, 2018: 1-23

[10] Song C, He K, Lin J, et al. Robust local features for improving the generalization of adversarial training//Proceedings of the International Conference on Learning Representations. Addis Ababa, Ethiopia, 2020

[11] Wong E, Rice L, Kolter J Z. Fast is better than free: Revisiting adversarial training//Proceedings of the International Conference on Learning Representations. Addis Ababa, Ethiopia, 2020

[12] Dong Y, Deng Z, Pang T, et al. Adversarial distributional training for robust deep learning. Advances in Neural Information Processing Systems, 2020, 33: 8270-8283

[13] Wang Y, Zou D, Yi J, et al. Improving adversarial robustness requires revisiting misclassified examples//Proceedings of the International Conference on Learning Representations. Addis Ababa, Ethiopia, 2020

[14] Jia X, Zhang Y, Wu B, et al. LAS-AT: Adversarial training with learnable attack strategy//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA, 2022: 13388-13398

[15] Xie C, Yuille A. Intriguing properties of adversarial training at scale//Proceedings of the International Conference on Learning Representations. Addis Ababa, Ethiopia, 2020

[16] Andriushchenko M, Flammarion N. Understanding and improving fast adversarial training. Advances in Neural Information Processing Systems, 2020, 33: 16048-16059

[17] Lecuyer M, Atlidakis V, Geambasu R, et al. Certified robustness to adversarial examples with differential privacy//Proceedings of the 2019 IEEE Symposium on Security and Privacy. San Francisco, USA, 2019: 656-672

[18] Cohen J, Rosenfeld E, Kolter Z. Certified adversarial robustness via randomized smoothing//Proceedings of the International Conference on Machine Learning. Los Angeles, USA, 2019: 1310-1320

[19] Lee G-H, Yuan Y, Chang S, et al. Tight certificates of adversarial robustness for randomly smoothed classifiers. Advances in Neural Information Processing Systems, 2019: 4911-4922

[20] Teng J, Lee G-H, Yuan Y. L1 Adversarial robustness certificates: A randomized smoothing approach//Proceedings of the International Conference on Learning Representations. Addis Ababa, Ethiopia, 2020

[21] Levine A, Feizi S. Robustness certificates for sparse adversarial attacks by randomized ablation//Proceedings of the AAAI Conference on Artificial Intelligence. 2020, 34(4): 4585-4593

[22] Chen R, Li J, Yan J, et al. Input-specific robustness certification for randomized smoothing//Proceedings of the AAAI Conference on Artificial Intelligence, Virtual, 2022: 6295-6303

[23] Salman H, Sun M, Yang G, et al. Denoised smoothing: A provable defense for pretrained classifiers. Advances in Neural Information Processing Systems, 2020, 33: 21945-21957

[24] Zhang Y, Yao Y, Jia J, et al. How to robustify black-box ML models? A zeroth-order optimization perspective//Proceedings of the International Conference on Learning Representations. Virtual, 2022

[25] Hong H, Wang B, Hong Y. UniCR: Universally approximated certified robustness via randomized smoothing//Proceedings of the Computer Vision. Aviv, Israel, 2022: 86-103

[26] Krizhevsky A, Hinton G. Learning multiple layers of features from tiny images [Ph. D. dissertation]. University of Tront, Tront, Canada, 2009

[27] Netzer Y, Wang T, Coates A, et al. Reading digits in natural images with unsupervised feature learning//Proceedings of the NIPS Workshop on Deep Learning and Unsupervised Feature Learning. Granada, Spain, 2011

[28] Zhang D, Ye M, Gong C, et al. Black-box certification with randomized smoothing: A functional optimization-based framework. Advances in Neural Information Processing Systems, 2020, 33: 2316-2626

[29] Fang G, Song J, Shen C, et al. Data-free adversarial distillation. Computing Research Repository, abs: 1912.11006, 2019

[30] Song L, Shokri R, Mittal P. Privacy risks of securing machine learning models against adversarial examples//Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security. Singapore, 2019: 241-257

LI Qiao, Ph. D. candidate. Her research area is artificial intelligence security.

CHEN Jing, Ph. D., professor, Ph. D. supervisor. His research interests include network security, distributed system security, and blockchain.

ZHANG Zi-Jun, Ph. D., associate professor, M. S. supervisor. His research areas is deep learning.

HE Kun, Ph. D., associate professor, M. S. supervisor. His research interests include applied cryptography, network

security, cloud computing security, artificial intelligence security, and blockchain security.

DU Rui-Ying, Ph. D., professor. Ph. D. supervisor. Her research interests include network security and privacy protection.

WANG Xin-Xin, Ph. D. candidate. Her research area is adversarial attack.

Background

The problem addressed in this paper belongs to the field of adversarial attacks and defense in Deep Neural Networks (DNNs). Adversarial attacks refer to the intentional manipulation of input data to deceive DNNs, causing them to produce incorrect outputs. Such attacks pose significant threats in critical domains like facial recognition and autonomous driving.

Various empirical and certified defense methods have been proposed to counter adversarial attacks on DNNs. Empirical defense methods aim to enhance model robustness through techniques such as adversarial training and detection. However, these methods often lack theoretical guarantees and are vulnerable to complex attacks. On the other hand, certified defense methods provide provable guarantees of DNN robustness. However, existing certified defense methods primarily rely on white-box assumptions, limiting their applicability in practical scenarios where model owners may be unwilling to share detailed information.

The objective of this paper is to propose a universal black-box certified defense method for DNNs, providing a robust defense solution for black-box models. The goal is to achieve robustness certification with only query access to the target model. The proposed method utilizes a query-based data-free substitute model generation approach, enabling robustness certification without requiring access to the internal details of the model. Furthermore, the method incorporates random smoothing and noise selection techniques, constructing a universal certified defense solution capable of resisting adversarial attacks involving any norm.

This research was supported in part by the National Key R&D Program of China under Grant No. 2022YFB3102100, the National Natural Science Foundation of China under Grant Nos. 62206203, 62076187, the Key R&D Program of Hubei Province under Grant No. 2022BAA039, the Key R&D Program of Shandong Province under Grant No. 2022CXPT055, and the Wuhan Science and Technology Program under Grant No. 2023010302020707.