

基于 softmax 激活变换的对抗防御方法

陈晋音^{1,2}, 吴长安², 郑海斌²

(1. 浙江工业大学网络空间安全研究院, 浙江 杭州 310023; 2. 浙江工业大学信息工程学院, 浙江 杭州 310023)

摘要: 深度学习广泛应用于图像处理、自然语言处理、网络挖掘等各个领域并取得良好效果, 但其容易受到对抗攻击、存在安全漏洞的问题引起广泛关注。目前已有一些有效的防御方法, 包括对抗训练、数据变化、模型增强等方法。但是, 依然存在一些问题, 如提前已知攻击方法与对抗样本才能实现有效防御、面向黑盒攻击的防御能力差、以牺牲部分正常样本的处理性能为代价、防御性能无法验证等。因此, 提出可验证的、对抗样本不依赖的防御方法是关键。提出了 softmax 激活变换防御 (SAT, softmax activation transformation), 这是一种针对黑盒攻击的轻量级的快速防御。SAT 不参与模型的训练, 在推理阶段对目标模型的输出概率进行隐私保护加固并重新激活, 通过 softmax 激活变换与深度模型防御的连接定义, 证明通过 softmax 函数的变换后能实现概率信息的隐私保护从而防御黑盒攻击。SAT 的实现不依赖对抗攻击方法与对抗样本, 不仅避免了制作大量对抗样本的负担, 也实现了攻击的事前防御。通过理论证明 SAT 的激活具有单调性, 从而保证其防御过程中正常样本的识别准确率。在激活过程中, 提出可变的 softmax 激活函数变换系数保护策略, 在给定范围内随机选择隐私保护变换系数实现动态防御。最重要的一点, SAT 是一种可验证的防御, 能够基于概率信息隐私保护和 softmax 激活变换推导其防御的有效性和可靠性。为了评估 SAT 的有效性, 在 MNIST、CIFAR10 和 ImageNet 数据集上进行了针对 9 种黑盒攻击的防御实验, 令所有攻击方法的平均攻击成功率从 87.06%降低为 5.94%, 与多种先进黑盒攻击防御方法比较, 验证了所提方法可以达到最优防御性能。

关键词: 深度学习; 对抗防御; 可验证; 攻击无关

中图分类号: TP181

文献标志码: A

DOI: 10.11959/j.issn.2096-109x.2022016

Novel defense based on softmax activation transformation

CHEN Jinyin^{1,2}, WU Changan², ZHENG Haibin²

1. Institute of Cyberspace Security, Zhejiang University of Technology, Hangzhou 310023, China

2. College of Information Engineering, Zhejiang University of Technology, Hangzhou 310023, China

Abstract: Deep learning is widely used in various fields such as image processing, natural language processing,

收稿日期: 2021-08-30; 修回日期: 2021-11-23

通信作者: 陈晋音, chenjinyin@zjut.edu.cn

基金项目: 国家自然科学基金 (62072406); 信息系统安全技术重点实验室基金 (61421110502)

Foundation Items: The National Natural Science Foundation of China (62072406), National Key Laboratory of Science and Technology on Information System Security (61421110502)

引用格式: 陈晋音, 吴长安, 郑海斌. 基于 softmax 激活变换的对抗防御方法[J]. 网络与信息安全学报, 2022, 8(2): 48-63.

Citation Format: CHEN J Y, WU C A, ZHENG H B. Novel defense based on softmax activation transformation[J]. Chinese Journal of Network and Information Security, 2022, 8(2): 48-63.

network mining and so on. However, it is vulnerable to malicious adversarial attacks and many defensive methods have been proposed accordingly. Most defense methods are attack-dependent and require defenders to generate massive adversarial examples in advance. The defense cost is high and it is difficult to resist black-box attacks. Some of these defenses even affect the recognition of normal examples. In addition, the current defense methods are mostly empirical, without certifiable theoretical support. Softmax activation transformation (SAT) was proposed in this paper, which was a light-weight and fast defense scheme against black-box attacks. SAT reactivates the output probability of the target model in the testing phase, and then it guarantees privacy of the probability information. As an attack-free defense, SAT not only avoids the burden of generating massive adversarial examples, but also realizes the advance defense of attacks. The activation of SAT is monotonic, so it will not affect the recognition of normal examples. During the activation process, a variable privacy protection transformation coefficient was designed to achieve dynamic defense. Above all, SAT is a certifiable defense that can derive the effectiveness and reliability of its defense based on softmax activation transformation. To evaluate the effectiveness of SAT, defense experiments against 9 attacks on MNIST, CIFAR10 and ImageNet datasets were conducted, and the average attack success rate was reduced from 87.06% to 5.94%.

Keywords: deep learning, adversarial defense, certifiable, attack-free

0 引言

深度学习具有强大的特征学习能力, 这为机器学习和人工智能的发展提供了巨大的机遇。因此, 深度学习在众多领域发挥了重要作用, 如图像识别^[1]、生物建模^[2]、疾病诊断^[3]、语音识别^[4]、自然语言处理^[5]等。尤其在计算机视觉领域, Krizhevsky 等^[6]展示了卷积神经网络在大规模视觉识别系统^[7]中的杰出性能, 因此深度学习成为计算机视觉领域关注的中心。随着深度神经网络模型的不断改进, 以及训练复杂模型所需硬件的持续升级, 深度学习得到广泛应用, 包括人脸识别^[8]、自动驾驶^[9]、视频监控^[10]、恶意软件检测^[11]、无人机和机器人控制^[12]等。

随着深度学习技术的深入应用, 其中隐藏的安全问题逐渐暴露。Szegedy 等^[13]率先发现了深度神经网络在图像分类中存在脆弱性——深度模型容易受到图像中细小扰动的对抗攻击, 这种细小扰动对人类视觉系统来说几乎是不可察觉的。处于自动驾驶模式的特斯拉汽车, 被道路上特定位置的几张贴纸欺骗, 自动并入反向车道。在这种安全要求高的应用中, 一旦深度模型暴露出漏洞, 对乘客和行人都是灾难性的后果。因此, 研究深度模型存在的漏洞并提供鲁棒的防御方法是关键。

为了发现深度模型中存在的安全漏洞, 学者

提出了一系列高效的对抗攻击方法。根据攻击者所能获取模型信息的程度, 对抗攻击可分为白盒攻击、灰盒攻击与黑盒攻击。其中, 白盒攻击是指攻击者可以获取目标模型的结构、参数、数据集、模型携带防御等所有信息后展开的攻击, 如快速梯度符号法^[14] (FGSM, fast gradient sign method)、MI-FGSM^[15]、DeepFool^[16]、C&W^[17]、基于生成式对抗网络的大规模对抗样本生成器^[18] (MAG-GAN, massive attack generator via GAN) 等; 灰盒攻击是指获取模型信息但不了解防御方法; 黑盒攻击则是在没有任何模型信息的前提下仅通过查询得到模型输出的类标或对应置信度, 利用优化计算或等价模型训练等方法实现攻击, 如 One-pixel^[19]、基于零阶优化的攻击^[20] (ZOO, zeroth order optimization)、POBA-GA^[21]等。白盒攻击大多存在于实验室的理论研究, 对于现在流行的预训练模型和提供 API 接口的商业模型, 黑盒攻击往往具有更大的威胁性^[22-23]。因此, 实际应用场景中, 黑盒攻击比白盒攻击更容易操作, 所以, 进行黑盒攻击防御研究更具现实意义。

黑盒攻击可分为基于输出置信度的攻击和基于决策边界的攻击^[24]。基于决策边界的攻击能获取到的信息更少, 攻击操作更困难。在攻击条件上, 基于决策的攻击需要对模型的访问次数是基于输出置信度攻击方法的几十倍, 甚至上百倍,

一定程度上,大量的重复性查询容易暴露,且攻击成功率相对较低。已经有的商用 AI 服务,包括谷歌的云服务 GCE、微软的云计算服务 Azure 等,它们均提供输出类标及对应置信度,容易受到基于输出置信度的黑盒攻击。与此同时,针对对抗攻击的防御方法研究如火如荼展开。根据防御实施的对象不同,可以分为数据修改的防御^[25]、模型修改的防御和附加模型的防御^[26]。根据防御操作是否需要对抗样本,可以分为攻击依赖的防御(如对抗训练、扰动蒸馏网络等),攻击无关的防御(如蒸馏防御、数据变换等)。攻击依赖的防御面对特定的对抗攻击具有很好的防御效果,但当攻击策略改变后,防御性能会受较大影响;攻击无关的防御在防御过程中不需要对抗样本支持,对不同攻击制作的对抗样本都能够防御,防御具有一定迁移性。

已有的防御方法在防御成功率上取得了不错的效果,但在实际应用中仍然会遇到以下问题。

1) 攻击依赖性:防御的效果依赖于先前生成的对抗样本,如对抗训练,而且遇到未知的攻击方法生成的对抗样本,则防御成功率受较大影响。

2) 防御代价大:防御的操作大多会影响正常样本的识别准确率,如数据变换随机操作虽然能够破坏对抗扰动,但也干扰了正常样本的识别。

3) 防御成本高:一次防御通常需要重新训练网络或者添加额外的子网络,增加原始模型的复杂性和冗余度,如对抗训练和附加网络防御需要占用大量的时空成本,无法实现轻量级的快速防御。

4) 防御滞后性:防御通常在攻击发生后采用,面对未知攻击,部分防御方法需要再次重新训练,无法实现事前防御。

针对这些问题,本文提出了 softmax 激活变换(SAT, softmax activation transformation)防御,这是一种针对黑盒攻击的轻量、快速的防御。本文方法通过对输出概率进行重激活实现防御,在不同的图片输入目标模型时,模型输出的概率分布基本相似,导致基于分数的黑盒攻击无效。SAT 不参与模型的训练,仅在推理阶段通过对目标模型的

输出概率进行重新激活,非常快速、轻量,解决了问题 3)。SAT 与攻击类型无关,不仅避免了制作大量对抗样本的负担,也实现了攻击的事前防御,解决了问题 1)和问题 4)。SAT 的激活具有单调性,因此不会影响正常样本的识别,解决了问题 2)。

本文的主要贡献包括以下 4 点。

1) 本文提出了一种针对黑盒攻击的快速防御方法,通过在模型的输出层引入 softmax 激活的变换保护机制,在推理阶段对模型输出进行重激活,具有可验证、低成本的防御优势。

2) SAT 防御方法不参与模型训练,不依赖于事前生成的对抗样本,对未知的黑盒攻击仍然具有防御效果,实现了攻击无关和事前防御。

3) 理论证明 SAT 的单调性,在实现对抗防御的同时不影响目标模型对正常样本的识别准确率,实现了防御代价最小化。

4) 本文在 MNIST、CIFAR10 和 ImageNet 数据集上进行了防御验证,对比现在最优防御方法,针对 9 种黑盒攻击的防御实验, SAT 将平均攻击成功率从 87.06%降低为 5.94%,达到最优防御性能。

1 相关工作

本节主要介绍实验中涉及的对攻击方法和对比的防御方法。

1.1 对抗攻击方法介绍

针对深度学习中存在的对抗攻击,根据对手掌握的目标模型信息的多少,可分为白盒攻击和黑盒攻击,具体区别如表 1 所示。由表 1 可知,对于面向公众的模型,研究针对黑盒攻击的模型防御能力是有现实意义的。

黑盒攻击分为基于分数和基于决策边界两种攻击方式^[24]。基于分数的黑盒攻击是指攻击者能够获取模型输出的概率,而在基于决策边界的黑盒攻击中,攻击者只能获取到模型输出的类标。基于决策边界的攻击能获取到的信息比基于分数的攻击更少,攻击操作更困难。在攻击条件上,基于决策的攻击需要对模型访问的次数是基于分数的攻击的几十倍,甚至上百倍,大量的重复性访问容易被目标模型所禁止,导致攻击失败。而且从攻击结果上看,即使访问次数不受模型限制,基于决策边界的攻击成功率仍然远低于基于分数

表1 白盒攻击和黑盒攻击的区别
Table 1 The difference between white-box and black-box attacks

描述	白盒攻击	黑盒攻击
对手可知信息	所有的信息：模型的结构、参数、训练时的数据集等	有限的信息：只能通过查询获得模型对探测性输入的输出结果
攻击策略	通过输入关于损失函数求导所获取的梯度信息生成对抗样本	通过等价模型、梯度估算、概率优化、决策选择等策略获取对抗样本
用途	用于实验室环境研究或者内部人员进行模型漏洞检测	对现流行的预训练模型或者提供 API 接口的商用模型进行信息窃取或性能破坏

的黑盒攻击，所以研究基于分数的黑盒攻击更具有意义。

基于分数的黑盒攻击可以分为梯度估计、模型等价、概率优化3种，以下一一阐述。

基于梯度估计的黑盒攻击：通过观察细微改变输入后分类概率的差异，利用零阶优化法直接估算出模型的梯度，然后沿着梯度方向优化输入，得到对抗样本。该方法首次体现在 Chen 等^[27]提出的 ZOO；Tu 等^[28]提出了 ZOO 的改进方法 Auto-ZOO，利用基于自编码的零阶优化算法大大减少了对目标模型的访问次数；Chen 等^[29]提出了 Boundary++，通过将决策边界攻击 Boundary^[24]与 ZOO 相结合的方法弥补了 Boundary 算法中模型访问次数过多的缺陷。

基于模型等价的黑盒攻击：攻击者训练一个替代模型，该替代模型与目标模型具有相似的分类边界，然后在等价模型上利用白盒攻击方法生成对抗样本，该对抗样本在目标模型上具有类似的攻击效果。该方法首先由 Papernot 等^[30]提出，使用基于雅可比矩阵的数据集增强来合成一个替代训练集。Zhang 等^[31]将这个想法扩展到基于 EEG 的 BCI，不同点在于使用输入计算的损失来合成一个新的训练集。

基于概率优化的黑盒攻击：在输入样本上直接加随机扰动，利用不同的优化策略最大化非原始正确类的概率来获取对抗样本。典型的方法有 Su 等^[32]提出的单像素点攻击(one-pixel)；Ilyas 等^[33]提出的基于自然进化策略攻击的 NES；Chen 等^[21]提出的基于进化计算的黑盒攻击 POBA-GA。

1.2 对抗防御方法介绍

根据防御对象的不同，可以将防御方法分为3种：基于数据修改的防御、基于模型修改的防御和基于附加网络的防御^[34]。根据防御方法是否

依赖于预生成的对抗样本，可分为攻击相关的防御和攻击无关的防御。

攻击相关的防御中，最为有效的是对抗训练，对抗训练首先由 Goodfellow 等^[35]提出，通过将攻击得到的对抗样本注入训练数据集，用来增强模型的鲁棒性；Kurakin^[36]和 Madry 等^[37]提出利用非迭代的快速梯度法产生的样本（如 FGSM）进行对抗训练，只会增强针对非迭代攻击的防御鲁棒性，而无法增强针对迭代攻击，如投影梯度下降法（PGD, project gradient descent）的鲁棒性。Madry 等^[38]提出了针对多步迭代 PGD（攻击进行对抗训练的方法），但与之之前的对抗训练相比，K-PGD 对抗训练大约需要多出 K+1 倍的计算量。为解决此问题，Shafahi 等^[39]提出了“免费”对抗训练。

攻击无关的防御，是指在不知道任何先验攻击的情况下执行防御操作。攻击相关防御对特定的攻击具有很好的防御效果，但对其他攻击，防御效果不佳；攻击无关防御通常对所有攻击（已有的和未知的）都具有相似的防御效果。简而言之，攻击相关防御具有局限性和滞后性的缺点，而攻击无关防御具有普适性和超前性的优点，所以攻击无关的防御更具有研究价值。典型的攻击无关防御包括：Xie 等^[40]提出的随机变换输入样本尺寸的防御；Guo 等^[41]通过对输入图像进行各种预处理操作提出了 Transform 防御；Strauss 等^[42]通过将各种深度网络集成提出了 Ensemble 防御；Liu 等^[43]通过在模型各层内注入随机噪声，实现了模型的自集成防御（RSE, random self-ensemble）。

2 深度模型和黑盒攻击的基本定义

2.1 黑盒模型的表示

正常情况下神经网络的前向传播过程可以表

示为 $f: \mathbb{R}^M \rightarrow \mathbb{R}^N$ ，其中 M 表示输入的维度， N 表示输出的维度。将样本 $x \in X$ 输入深度模型中进行分类预测，在经过最后一层全连接后得到一个向量 $\mathbf{Z}(x, i)$ ，其中 $i \in \{0, 1, 2, \dots, C-1\}$ ， C 表示数据集的分类总数，向量 $\mathbf{Z}(x, i)$ 表示输入 x 被分类成第 i 类时的权重值，即得分情况，该向量被称为目标模型的逻辑输出 (logits)。为了将每一类的 logits 值归一化，使用 softmax 函数对其进行激活，得到包含各类 logits 值归一化后的概率向量 $\mathbf{f}(x)$ ，向量 $\mathbf{f}(x)$ 表示输入 x 被分类成各类时的概率大小，概率值最大的那一类即模型分类结果。logits 向量 $\mathbf{Z}(x, i)$ 通过 softmax 激活函数转化为概率向量 $\mathbf{f}(x)$ 的计算方式如下。

$$f^{(i)}(x) = \frac{e^{\mathbf{Z}(x, i)}}{\sum_{i=0}^{C-1} e^{\mathbf{Z}(x, i)}} \quad (1)$$

其中， e 表示自然底数， x 为目标模型的输入， $f^{(i)}(x)$ 表示输入被分类为第 i 类的概率，则目标模型的输出类标表示为 $\hat{y} = \arg \max(\mathbf{f}(x))$ ，其中 $\hat{y}$ 表示 x 的预测类标， $\arg \max(\cdot)$ 返回数值元素值最大的位置的坐标。给定 x 的真实标签为 y ，当 $\hat{y} = y$ 时，则预测正确，反之则预测错误。

2.2 黑盒攻击的表示

在黑盒对抗攻击中，攻击者通常能获取模型各类的输出概率 $\mathbf{f}(x)$ ，然后利用不断改变的输入样本 x ，查询模型的输出，以达到攻击的目的。以图像分类任务为例，具体实现定义为

$$\begin{aligned} \hat{y}^* &= \arg \max(\mathbf{f}(x^*)) \\ \text{s.t.} \quad &\|x - x^*\|_p \leq \rho \\ &x, x^* \in [0, 1]^{c \cdot w \cdot h} \end{aligned} \quad (2)$$

其中， x^* 表示对抗样本， $\hat{y}^*$ 表示对抗样本的预测类标， ρ 表示可允许的对抗扰动尺寸， c 表示输入图像的通道， w 表示宽度， h 表示高度， $\|\cdot\|_p$ 表示计算 L_p 范数。当 $\hat{y} = y$ 且 $\hat{y}^* \neq y$ 时，则对抗样本攻击成功。

本文主要考虑基于分数的黑盒攻击。基于决策的黑盒攻击的查询次数远远高于基于分数的黑盒攻击，且攻击得到的对抗样本扰动一般较大，对基于分数的黑盒攻击实现防御更加具有现实意义。基于分数的黑盒攻击主要包括梯度估计、模型等价、概率优化 3 种，具体介绍如下。

1) 基于梯度估计的黑盒攻击。通过对细微修改的输入进行重复查询并记录返回值的细微差异，直接估计目标模型的梯度。梯度估计公式如下。

$$\hat{g}(x') := \frac{\partial f(x')}{\partial x'} = \lim_{h \rightarrow 0} \frac{f(x' + h\mathbf{e}) - f(x' - h\mathbf{e})}{2h\mathbf{e}} \approx \frac{f(x' + h\mathbf{e}) - f(x' - h\mathbf{e})}{2h} \quad (3)$$

其中， h 是一个很小的常数， $\mathbf{e}$ 是标准基向量， x' 表示在生成对抗样本的迭代过程中的中间量， $\hat{g}(\cdot)$ 表示估计梯度。最终的对抗样本为 $x^* = x + \xi \cdot \hat{g}(x)$ ，其中 ξ 表示扰动步长。

2) 基于模型等价的黑盒攻击。攻击者首先通过训练一个与目标模型分类边界相似的替代模型，表示为 $f'(\cdot)$ ，进一步对替代模型进行白盒攻击得到对抗样本；然后利用得到的对抗样本攻击目标模型。替代模型和目标模型十分相似，因此能够攻击替代模型的对抗样本也能攻击目标模型。替代模型的梯度表示为

$$g'(x') := \frac{\partial f'(x')}{\partial x'} \quad (4)$$

其中， x' 表示在生成对抗样本的迭代过程中的中间量， $g'(\cdot)$ 表示替代模型 $f'(\cdot)$ 的梯度。最终的对抗样本为 $x^* = x + \xi \cdot g'(x)$ ，其中 ξ 表示扰动步长。

3) 基于概率优化的黑盒攻击。攻击者利用目标模型的输出概率 $\mathbf{f}(x)$ 作为优化目标，通过判定查询前后优化概率的大小，确定下一步的搜索方向。在限制对抗扰动尺寸的情况下，最小化目标类的输出概率，优化目标如下。

$$\begin{aligned} \underset{\rho_r}{\text{minimize}} \quad &\{f^{(i)}(x + \rho_r)\} \\ \text{s.t.} \quad &\|\rho_r\|_p \leq \rho \end{aligned} \quad (5)$$

其中， $\text{maximize}\{\cdot\}$ 表示通过优化 ρ_r 实现 $f(\cdot)$ 在真实类 i 上的概率最小化， ρ_r 的 L_p 范数受限 ρ 。

3 基于激活函数变换的动态防御方法

本节详细介绍基于 softmax 激活函数变换的对抗防御方法，包括基于概率信息隐私保护防御验证、基于 softmax 激活函数变换的对抗防御方法 SAT、基于可变变换系数实现动态防御，进一步证明 SAT 在保持良性样本识别的有效性，并分析 SAT 的算法复杂度。

SAT 防御方法的本质是在目标模型的推理阶段进行 softmax 激活函数变换，通过令任意样本的任意标签的分类概率分布接近，降低模型输出概率信息的泄露风险。以 CIFAR10 数据集的分类任务，基于 ResNet50 结构的目标模型为例，基于 softmax 激活变换的对抗防御方法 SAT 框架如图 1 所示。防御前，模型输出的不同类之间的概率差异十分明显，黑盒攻击方法能够通过梯度估计、模型等价、概率优化等操作实现对抗样本的生成，

即图 1 中的“clean”样本经过模型输出概率的反馈生成“adversary”。防御后，原来的输出概率经过 SAT 激活，从而对不同类之间的概率信息进行脱敏，无法反馈给黑盒攻击方法生成对抗样本，即攻击失败，实现对抗防御。进一步，借助 t -SNE 降维方法对防御前后的概率分布进行可视化，直观地解释了 SAT 针对黑盒对抗攻击的防御有效性，是通过增加概率分布的混乱程度实现的。

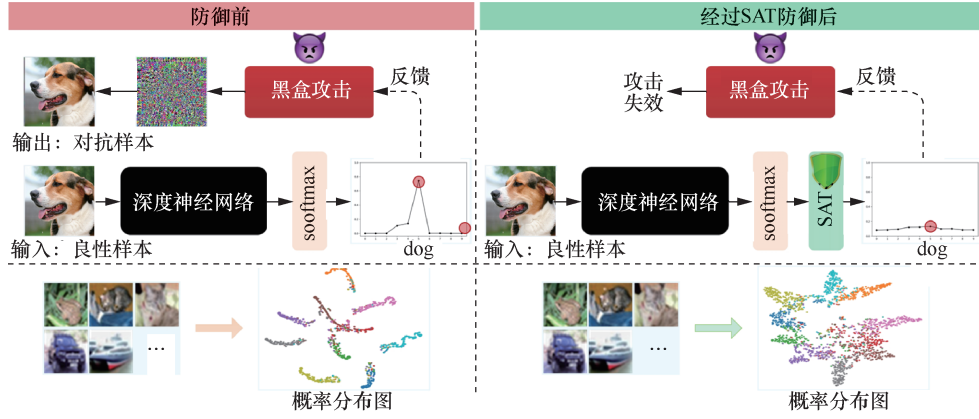


图 1 基于 softmax 激活变换的对抗防御方法 SAT 框架
Figure 1 The framework of adversarial defense based on activation transformation

SAT 不改变目标模型的训练操作，因此具有攻击无关、针对未知攻击进行事前防御的优点；SAT 的激活是单调的，因此具有不影响正常样本识别的优点；SAT 只需要修改推理阶段的激活函数，因此具有轻量、低成本、快速的防御优点。下面对 SAT 的具体操作进行详细介绍。

3.1 基于概率信息隐私保护的防御验证

3.1.1 目标模型概率信息保护的指数机制

给定一个数据样本 x 和一组可能的输出集合 Y ，概率信息保护的指数机制选择的结果满足以下概率性质。

$$f^{(i)}(x) \propto e^{(\varepsilon \times f^{(i)}(x)) / (2 \times s(f, \|\cdot\|))} \quad (6)$$

其中， $Y = \{0, 1, 2, \dots, C-1\}$ 表示给定输入数据样本 x 在该机制中的合法输出集合，即数据集的类标集合； i 表示输入数据样本 x 的合法输出，即样本 x 的真实类标； $f^{(i)}(x)$ 表示输入为数据样本 x ，输出为 i 时的得分函数，即目标模型的输出概率； e 表示自然底数， ε 表示隐私保护变换系数； $s(\cdot, \cdot)$ 表示得分函数 $f(\cdot)$ 的敏感度，定义为

$$s(f, \|\cdot\|) = \max_{\|x_i, x_j\|_p \leq \rho} \|f^{(i)}(x) - f^{(i)}(x')\| \quad (7)$$

其中， x_i, x_j 表示两个不同的样本，计算敏感度时，对数据样本 x 和 x' 的距离条件进行松弛， $\{x, x'\} \in X$ 。注意，式(7)中仅用到 i 类，即样本 x 和 x' 被预测为 i 类的概率差值， x' 的真实类标是 k ，且 $k \neq i$ 。对于给定的 x, x' 可以是任意样本，因此 i 和 j 是不相关的。

3.1.2 针对基于梯度估计的黑盒攻击的防御验证

在基于梯度的黑盒攻击中，对式(3)进行展开，得到

$$\begin{aligned} f(x' + \text{he}) &= \{f^{(0)}(x' + \text{he}), f^{(1)}(x' + \text{he}), \dots, f^{(C-1)}(x' + \text{he})\} \\ f(x' - \text{he}) &= \{f^{(0)}(x' - \text{he}), f^{(1)}(x' - \text{he}), \dots, f^{(C-1)}(x' - \text{he})\} \end{aligned}$$

根据式(6)，当 $\varepsilon \rightarrow 0$ 时，有

$$\lim_{\varepsilon \rightarrow 0} (f^{(0)}(x' + \text{he}) - f^{(0)}(x' - \text{he})) \approx 0$$

$$\lim_{\varepsilon \rightarrow 0} (f^{(1)}(x' + \text{he}) - f^{(1)}(x' - \text{he})) \approx 0$$

...

$$\lim_{\varepsilon \rightarrow 0} (f^{(C-1)}(x' + \text{he}) - f^{(C-1)}(x' - \text{he})) \approx 0$$

则式(3)估计梯度变为

$$\hat{g}(x') := \frac{\partial f(x')}{\partial x'} \approx \frac{f(x' + h\mathbf{e}) - f(x' - h\mathbf{e})}{2h}$$

由此可得, SAT 令 $f(x' + h\mathbf{e})$ 和 $f(x' - h\mathbf{e})$ 的差异变得足够小, 此时攻击者无法对梯度 $\hat{g}(x')$ 进行估计, 导致无法实现对抗样本生成, 能够使这一类攻击失效。

3.1.3 针对基于模型等价的黑盒攻击的防御验证

在基于模型等价的黑盒攻击中, 其本质是利用对抗样本在相似模型间的攻击迁移性, 因此替代模型和目标模型之间的分类边界拟合程度直接决定了对抗样本的黑盒攻击效果。在替代模型的生成过程中, 使用目标模型的输出概率作为真实概率值, 实现替代模型的输出概率分布逼近目标模型。基于交叉熵的优化目标定义为

$$\text{Loss}_{\text{CE}} = -\frac{1}{m} \sum_{k=1}^m \sum_{i=0}^{C-1} (f^{(i)}(x_k) \log(f^{(i)}(x_k))) \quad (8)$$

其中, m 表示样本个数, x_k 表示数据集中第 k 个样本, $\log(\cdot)$ 表示对数函数, $f^{(i)}(\cdot)$ 和 $f^{(i)}(\cdot)$ 分别表示目标模型和替代模型预测 x_k 属于第 i 类的输出概率。最小化 Loss_{CE} , 能够实现替代模型的输出概率分布接近目标模型的输出概率分布。

当 $\varepsilon \rightarrow 0$ 时, 进一步松弛样本 x 和 x' 的距离, 二者属于不同的两个类。为了保护输出概率的隐私安全, 输出概率需要同时满足以下两个条件。

$$f^{(i)}(x) > e^\varepsilon f^{(i)}(x') \quad (9)$$

$$f^{(i)}(x) > \max_{k:k \neq i} f^{(k)}(x) \quad (10)$$

其中, i 和 k 分别是样本 x 和 x' 的真实类标, $i, k \in Y$ 且 $i \neq k$, 则有

$$\begin{aligned} f^{(i)}(x) &\leq e^\varepsilon \cdot f^{(i)}(x'), & f^{(i)}(x) &> f^{(k)}(x) \\ f^{(k)}(x) &\leq e^\varepsilon \cdot f^{(k)}(x'), & f^{(k)}(x') &> f^{(i)}(x') \end{aligned}$$

可得

$$\lim_{\varepsilon \rightarrow 0} (f^{(i)}(x)) \approx \lim_{\varepsilon \rightarrow 0} (f^{(k)}(x)) \quad (11)$$

即同一个样本的不同类之间的输出概率十分接近。根据式(11)对式(8)进行变换, 可得

$$\text{Loss}_{\text{CE}} = -\frac{1}{m} \sum_{k=1}^m \sum_{j \in Y, j \neq i} (f^{(i)}(x_k) \log(f^{(i)}(x_k))) \quad (12)$$

即替代模型对目标模型的输出概率的学习是混乱的。对于属于第 i 类的样本, 学习第 i 类的输出概率值与学习第 j 类的输出概率值是没有差异的。

由此可得, SAT 通过输出模型概率信息保护的指数机制保护目标模型的输出概率, 使替代模型的等价过程出错, 能够使这一类攻击失效。

3.1.4 针对基于概率优化的黑盒攻击的防御验证

假设扰动搜索函数为 $\text{Search}(x, \min(f^{(i)}(x)))$, 表示向着最小化目标概率 $f^{(i)}(x)$ 的方向优化, 返回搜索方向 $\text{dire} \leftarrow \text{Search}(\cdot, \cdot)$ 。根据搜索方向更新输入样本 $x_{j+1} = x_j + \text{dire} \cdot \xi$, 其中 x_j 和 x_{j+1} 表示迭代前后的样本, ξ 表示迭代步长。

若满足 $f^{(i)}(x_{j+1}) < f^{(i)}(x_j)$, 即更新后的样本输出的真实类标的概率小于更新前的样本, 则继续往 dire 搜索, 反之更新搜索方向。

为避免 $\text{Search}(\cdot, \cdot)$ 函数找到 dire , SAT 使用可变的隐私保护变换系数。根据 3.1.1 节可得, 目标模型真实类的输出概率正比于 $e^{(\varepsilon \cdot f^{(i)}(x)) / (2s(f, \|\cdot\|))}$ 。若 ε 改变, 则 $f^{(i)}(x)$ 改变; $\text{Search}(\cdot, \cdot)$ 无法知道是更改 ε 引起的 $f^{(i)}(x)$ 变小, 还是正确的搜索方向 dire 引起的 $f^{(i)}(x)$ 变小。

对于任意的 $\varepsilon > 0$, ε_j 和 ε_{j+1} 分别表示第 j 次和第 $j+1$ 次迭代使用的隐私保护变换系数。存在这样的情况

$$\begin{cases} f^{(i)}(x_{j+1}) > f^{(i)}(x_j), & \varepsilon_j = \varepsilon_{j+1} \\ f^{(i)}(x_{j+1}) < f^{(i)}(x_j), & \exists \varepsilon_j, \varepsilon_{j+1} \end{cases} \quad (13)$$

若 $\text{Search}(\cdot, \cdot)$ 错误地将此时的 dire 作为搜索方向, 将会使攻击向错误的方向优化。

由此可得, SAT 通过可变隐私保护变换系数策略, 实现 $f(\cdot)$ 输出的真实类的最高概率值动态变化, 令黑盒攻击向错误的方向搜索扰动, 从而使这一类攻击失效。

3.2 基于概率信息保护的 softmax 激活变换

本文基于模型输出概率信息保护的指数机制, 设计在目标模型推理阶段的 softmax 激活变换, 实现模型输出概率的隐私保护, 达到抵御黑盒攻击的目的, 具体实施方案如下。首先, 重新定义在对抗攻防中实现概率信息保护的指数机制。

3.2.1 基于概率信息保护的 softmax 激活变换

进一步设计 softmax 激活函数 SAT, 具体公

式定义如下。

$$f_{\text{SAT}}^{(i)}(x) = \frac{e^{(\varepsilon \cdot f^{(i)}(x))/(2s(f, \|\cdot\|))}}{\sum_{j=0}^{C-1} e^{(\varepsilon \cdot f^{(j)}(x))/(2s(f, \|\cdot\|))}} \quad (14)$$

其中, $f_{\text{SAT}}(x_i)$ 激活函数的输入是 softmax 函数的输出, 最后的预测类标为 $\hat{y} = \arg \max(f_{\text{SAT}}(x))$ 。

如图 1 所示, 经过 softmax 函数激活后, 推理样本能够被正确识别为“dog”类, 且对应类的概率激活十分明显, 容易被黑盒攻击利用。而使用 SAT 激活后, 推理样本仍然能够被正确识别为“dog”, 但对应类的概率激活趋势十分平坦, 能够保护概率信息避免被黑盒攻击利用。

3.2.2 攻击无关性和防御快速性分析

攻击无关性分析: 根据防御的具体步骤可知, SAT 不参与目标模型的训练, 不需要使用对抗样本的任何信息, 能够实现攻击无关的事前防御, 对用户十分友好。

防御快速性分析: SAT 在推理阶段对 softmax 函数的输出进行重新激活, 输入为 $f(x)$, 涉及的变量包括隐私保护变换系数 ε 、敏感度 s 。在确定变量后, SAT 的计算十分方便, 能够实现轻量快速的防御。

3.2.3 基于局部敏感度的概率信息保护

SAT 采用的是局部敏感度策略, 令 $u = f^{(i)}(x) \in (1/C, 1]$, $u' = f^{(i)}(x') \in [0, 1]$, 则将式(6)变换为如下形式。

$$s(f, \|\cdot\|) = \max\{u, 1-u\} \quad (15)$$

其中, x' 可以是 X 中的任意样本, 因此存在 $\max\{u'\} = 1$, $\min\{u'\} = 0$ 。进一步, 敏感度可以表示为

$$s(f, \|\cdot\|) = \begin{cases} u, & u \geq 0.5, \exists u' = 0 \\ 1-u, & u < 0.5, \exists u' = 1 \end{cases} \quad (16)$$

得到更新后的概率条件为

$$f^{(i)}(x) \propto \begin{cases} e^{\varepsilon/2} & u \geq 0.5, \exists u' = 0 \\ e^{(\varepsilon \cdot u)/(2(1-u))} & u < 0.5, \exists u' = 1 \end{cases} \quad (17)$$

使用不同敏感度策略的概率条件分布如图 2 所示。可以看出, 随着输入概率 u 的变化, 全

局敏感度策略下的概率条件分布一直上升, 上升斜率可以认为基本不变; 而局部敏感度策略下的条件概率先上升, 且上升速度比全局敏感度策略下更快, 达到拐点后保持不变。相比全局敏感度, 局部敏感度前半段快速上升能够更好地进行正常样本的识别, 后半段的保持不变能够更好地避免黑盒攻击利用多次查询的概率泄露问题。

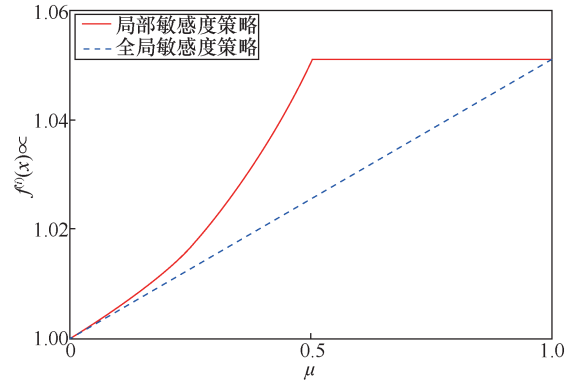


图2 不同敏感度策略的概率条件分布
Figure 2 Probability condition distribution of different sensitivity strategies

3.3 可变隐私保护变换系数的动态防御

隐私保护变换系数 ε 衡量隐私信息的泄露程度, ε 越小则隐私保护效果越好, 但是也会影响原始任务的正常工作。图 3 直观地显示了不同隐私保护变换系数下的概率分布, ε 越小, 概率分布越趋于平坦, 信息泄露越少。

不同隐私保护变换系数下的输出概率趋势如图 4 所示。以 CIFAR10 数据集在 ResNet50 模型上的输出概率保护为例, 图 4 展示了单张样本在不同 ε 下的概率信息泄露。当 ε 变小, 输出概率逐渐趋于平缓, 等价模型和梯度估计等黑盒攻击方式将会失效。为了保证模型的正常工作, 需要令真实类标所在位置的概率值一直保持最大, 这就导致基于概率优化的黑盒攻击方法在足够大量的查询次数后, 实现有效攻击。

为了令基于概率优化的黑盒攻击方法失效, SAT 引入了可变隐私保护变换系数策略, 即使用变化的 ε 实现动态的防御。直观地, 对于同一张输入样本, 真实类的概率值在每次查询时是变化的, 如式(13)所示, 错误的搜索方向增加了真实类的输出概率, 变化的 ε 降低了真实类的输出概

率, 且后者的改变程度大于前者, 则攻击方法误认为是正确的搜索方向降低了真实类的输出概率, 继续往错误的方向搜索, 从而使基于概率优化的攻击失效。

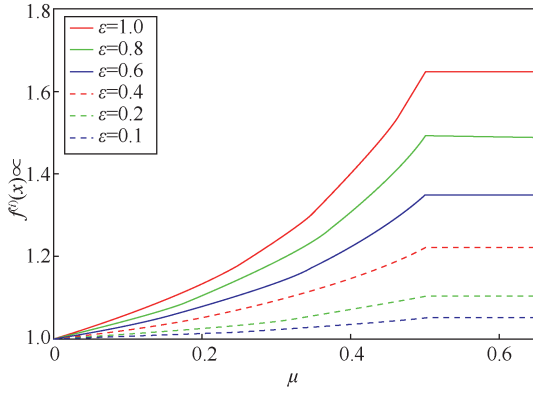


图 3 不同隐私保护变换系数下的概率条件分布
Figure 3 Probability condition distribution under different privacy protection budgets

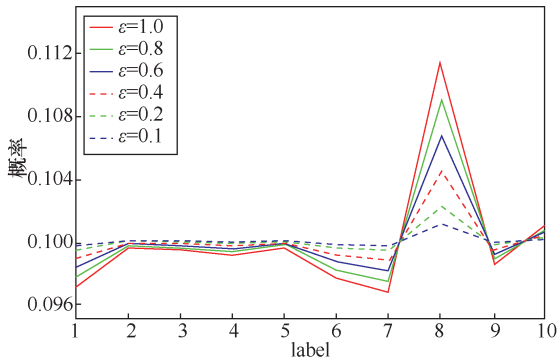


图 4 不同隐私保护变换系数下的输出概率趋势
Figure 4 Trend of output probabilities under different privacy protection budgets

3.4 SAT 保持正常样本识别的单调性证明

SAT 虽然降低了真实类的输出概率, 但是始终保持真实类的概率高于其他类的概率, 这是 $f_{\text{SAT}}(\cdot)$ 的单调性决定的。由式(17)可知, 当 $u \geq 0.5$ 时, $f^{(i)}(x) \propto e^{\varepsilon/2}$, 此时的 $e^{\varepsilon/2}$ 是一个常数, 因此 $f^{(i)}(x)$ 的斜率为 0。当 $u < 0.5$ 时, $f^{(i)}(x) \propto e^{(\varepsilon u)/(2(1-u))} = e^{\frac{\varepsilon}{2/u-2}}$ 。此时随着 u 增加, $e^{\frac{\varepsilon}{2/u-2}}$ 增加, $f^{(i)}(x)$ 单调递增。

因此, 在输入自变量 $u \in (1/C, 1]$ 的定义域内, 可以认为 $f^{(i)}(x)$ 是单调递增的, 能够保持 softmax 激活中概率最大的值在 SAT 激活中也是最大, 即不会改变原始的预测类标。因此以下结论成立: SAT 在降低目标模型概率信息泄露的同时, 能够

保证原始任务正常工作。

3.5 SAT 的防御复杂度分析

防御方法的复杂度分析包括训练阶段的复杂度和推理阶段的复杂度分析。从图 1 可以看出, SAT 只在推理阶段工作, 训练阶段的时间复杂度和空间复杂度都是 0。根据式(14)可知, SAT 防御是在原始的推理过程中增加了一步计算, 增加的存储空间是输入的 softmax 函数激活, 与分类数有关。因此, 在推理阶段, SAT 增加的空间复杂度是 $O(1)$, 增加的时间复杂度也是 $O(1)$ 。

4 实验与结果

本节首先介绍实验的基本设置, 包括软硬件环境、数据集、模型、攻击方法、防御方法和度量指标; 之后, 将本文提出的防御方法与其他方法在 9 种黑盒攻击方法上进行防御效果的对比, 并比较了各个防御方法的防御代价; 最后, 对实验的防御有效性、参数敏感性进行了分析。

4.1 实验设置

实验硬件及软件平台: i7-7700K 4.20GHz x8 (CPU), TITAN Xp 12GiBx2 (GPU), 16GB x4 memory (DDR4), Ubuntu16.04 (OS), Python(3.6), tensorflow-gpu (1.12.0), keras (2.2.4), torch (0.4.1) 和 torchvision (0.2.1)。

数据集: 实验采用 MNIST、CIFAR10 和 ImageNet 这 3 个公共数据集。其中, MNIST 数据集包括 60 000 张训练样本和 10 000 张测试样本, 样本大小是 28×28 的灰度图像, 共 10 类; CIFAR10 数据集由 50 000 张训练样本和 10 000 张测试样本组成, 样本是大小为 $32 \times 32 \times 3$ 的彩色图片, 共 10 类; ImageNet 数据集选取 130 万张训练样本和 5 万张测试样本共 1 000 类进行实验。

深度模型: 3 个数据集的默认模型均为 ResNet50, 等价模型均为 VGG19, 模型集成防御由 ResNet50、AlexNet 和 VGG19 组成。

攻击方法: 梯度估计为 ZOO、Auto-ZOO 和 Boundary++; 等价模型为基于雅克比的模型等价方法, 等价模型上的白盒攻击为 FGSM、MI-FGSM 和 C&W; 概率优化为 One-pixel、NES(PI)、POBA-GA。

防御方法: 本文采取 4 种防御方法作为对比

算法, 包括对抗训练 (AdvTrain)、通道变换 (Transform)、模型集成 (Ensemble) 和模型自集成 (RSE)。

AdvTrain: 在模型训练阶段, 将 FGSM 和 PGD 的对抗样本和正常样本混合在一起作为模型的训练集, 并重新训练得到防御模型。

Transform: 在模型的训练和推理阶段, 加入样本变换的预处理操作, 操作包括图像压缩和方差最小化操作。

Ensemble: 在模型训练和推理阶段, 将模型 ResNet50、AlexNet 和 VGG19 的输出层进行集合, 得到融合了 3 个模型的 Ensemble 模型。

RSE: 在模型训练和推理阶段, 均对模型的各层中加入随机噪声, 扰乱模型的梯度信息。

评价指标: 本文在验证防御有效性时, 采取防御超前性的方法, 即对模型先防御后攻击的形式, 而非先攻击再利用防御模型重新识别对抗样本, 前者更能表现防御模型在面对攻击时的防御效果。实验的评价指标是攻击成功率而非防御成功率。攻击成功率公式如下。

$$ASR = \frac{n_{adv}}{n_{right}} \quad (18)$$

其中, n_{right} 表示攻击前模型分类正确的样本数, n_{adv} 表示模型分类正确的样本中攻击成功的对抗样本数。攻击成功率越低, 表示防御效果越好。

4.2 基于梯度估计的黑盒攻击防御效果对比

本节实验主要分析 SAT 在基于梯度估计的黑盒攻击上同其他防御方法的防御效果对比。本文在 MNIST、CIFAR10 和 ImageNet 这 3 个数据集上利用基于梯度估计的 3 种黑盒攻击 (ZOO、Auto-ZOO 和 Boundary++) 直接对 6 个模型进行攻击, 模型分别为不采取防御的原始模型、AdvTrain、Transform、Ensemble 和 RSE。从实验结果可知, 在面对基于梯度估计的不同黑盒攻击方法时, SAT 相比其他防御方法具有更强的鲁棒性。具体实验结果如表 2 所示。

由表 2 的第二行可得, 每一种攻击方法下的括号中都对应一个数字, 该数字代表每种攻击方法在攻击不同策略的防御模型时所允许的最大模型访问次数, 如第 2 列中的“ZOO”对应的“(10 000)”,

是指利用 ZOO 攻击每一种防御模型时, 最多允许访问防御模型 10 000 次, 超出则视为攻击失败。由于不同攻击方法的攻击性能不一致, 实验为了保证每种攻击方法在原始模型上都具有较高的攻击成功率, 设定每一个数据集下的每一种攻击方法都对应不同的模型访问次数, 该访问次数由对应论文中涉及的访问次数和本次实验中在不防御情况下攻击成功率达到一定指标后的访问次数共同决定。基于梯度估计的黑盒攻击的成功率与允许的模型访问次数有很大关联, 通常攻击成功率与被允许的访问次数成正比, 所以在实验中, 相同攻击方法攻击不同防御模型时设定一样的访问次数, 以此更加客观真实地反映不同防御方法的防御效果。

由表 2 可知, 仅仅利用模型输出的概率, 3 个数据集下的 3 种攻击方法在攻击原始模型时, 攻击成功率都能达到 80% 以上, 甚至在两个小数据集上几乎能达到 100%。由此可知, 不采取防御措施的原始模型在面对基于梯度估计的黑盒攻击时是脆弱的, 防御措施必不可少。在攻击防御模型时, 攻击成功率大幅度下降, 但是对抗训练的防御效果明显弱于其他攻击方法, 这是因为对抗训练是攻击相关性防御, 只对特定的攻击有很好的防御作用, 其他攻击防御效果不佳。此外, RSE 的防御效果优于一般的 Ensemble 防御, 这是因为在 RSE 中不管是训练阶段还是推理阶段都加了随机扰动, 具有随机性的防御模型往往比一般的固定性防御更难被攻击, 因为攻击过程中往往会被随机性引到错误的方向, 但具有随机性的模型通常会以降低准确率为代价。由表 2 最后一行可得, 在大多数情况下, SAT 的防御效果相比其他防御方法是最优的。

4.3 基于模型等价的白盒攻击防御效果对比

本节实验主要分析 SAT 在基于模型等价的白盒攻击上同其他防御方法的防御效果对比。本文通过基于雅克比的模型等价方法用 VGG19 模型分别拟合 6 组模型的分界, 得到对应的 6 个等价模型, 再利用 3 种白盒攻击方法 (FGSM、MI-FGSM 和 C&W) 攻击这 6 个等价模型获取不同的对抗样本, 其中 FGSM 是通过计算单步梯度快速生成对抗样本的方法, 在白盒环境下, 通过

表 2 基于梯度估计的黑盒攻击防御效果对比
Table 2 Comparison of black box attack defense effects based on gradient estimation

模型	MNIST (ASR)			CIFAR10 (ASR)			ImageNet (ASR)		
	ZOO (10 000)	Auto-ZOO (1 000)	Boundary++ (5 000)	ZOO (5 000)	Auto-ZOO (1 000)	Boundary++ (5 000)	ZOO (250 000)	Auto-ZOO (5 000)	Boundary++ (80 000)
原始模型	99.57%	100.00%	100.00%	97.42%	100.00%	100.00%	89.31%	100.00%	84.54%
AdvTrain	28.56%	30.18%	35.73%	27.76%	28.10%	30.12%	21.36%	25.54%	24.79%
Transform	23.52%	24.56%	29.25%	23.93%	24.43%	22.65%	14.31%	16.96%	15.62%
Ensemble	20.49%	25.63%	19.54%	23.15%	20.53%	17.53%	14.49%	13.84%	10.89%
RSE	13.43%	14.15%	9.43%	8.59%	12.61%	10.75%	8.31%	9.62%	3.63%
SAT	6.42%	9.74%	8.83%	7.93%	8.94%	7.53%	7.65%	8.32%	6.38%

求出模型对输入的导数, 然后用符号函数得到其具体的梯度方向, 接着乘以一个步长, 得到的“扰动”加在原来的输入上就得到了在 FGSM 攻击下的样本; MI-FGSM 是一种基于动量迭代梯度的对抗攻击方法, 用梯度迭代扰动输入, 使损失函数最大化, 提高生成的对抗性样本的成功率; C&W 攻击是一种基于迭代优化的低扰动对抗样本生成方法, 通过限制 l_∞ 、 l_2 和 l_0 范数使扰动无法被察觉。最后, 利用这些对抗样本对 6 个被等价的目标模型进行攻击。从实验结果可知, 相比其他 4 个防御模型, SAT 明显具有更好的防御效果。具体实验结果如表 3 所示。

由表 3 可知, 在不防御的原始模型中, 基于模型等价的黑盒攻击的攻击效果比梯度估计的攻击效果差, 这是因为模型等价攻击依赖于对抗样本的迁移性, 即使等价模型和原模型的分界边界几乎一致, 但其结构以及参数上的差异性仍然会影响对抗样本的迁移效果。通过比较不同攻击方

法的攻击成功率, 可以观察到, MI-FGSM 的对抗样本的迁移性明显优于其他两种方法, 这是因为 MI-FGSM 利用动量迭代的方式生成对抗样本, 大大提升了对抗样本的迁移性。此外, AdvTrain 对 FGSM 的防御效果明显优于其他两种攻击方法, 这是因为对抗训练的训练集中包含了 FGSM 在 ResNet50 上的对抗样本。最后, 由表 3 最后一行可知, SAT 在等价模型上的防御效果远优于其他防御模型, 因为 SAT 虚假输出了原始模型的分界边界, 等价模型拟合的是扰乱后的边界, 并非模型真正的分类边界, 所以由该等价模型生成的对抗样本的迁移性很弱。

4.4 基于概率优化的黑盒攻击防御效果对比

本节实验主要分析 SAT 在基于概率优化的黑盒攻击上同其他防御方法的防御效果对比。本节实验的操作形式类似于 4.2 节, 只是 3 种攻击方法不同, 本节实验在 MNIST、CIFAR10 和 ImageNet 这 3 个数据集上利用基于概率优化的 3

表 3 基于模型等价的黑盒攻击防御效果对比
Table 3 Comparison of black box attack defense effects based on model equivalence

模型	MNIST (ASR)			CIFAR10 (ASR)			ImageNet (ASR)		
	FGSM	MI-FGSM	C&W	FGSM	MI-FGSM	C&W	FGSM	MI-FGSM	C&W
原始模型	75.83%	84.38%	72.75%	73.27%	86.84%	70.53%	69.94%	80.72%	67.52%
AdvTrain	23.57%	29.47%	35.38%	21.52%	29.99%	35.41%	16.10%	34.27%	30.63%
Transform	37.84%	44.42%	35.45%	33.40%	46.13%	31.14%	30.92%	43.71%	22.64%
Ensemble	30.45%	39.72%	28.42%	31.48%	40.25%	32.29%	25.55%	38.12%	23.41%
RSE	18.83%	28.31%	16.37%	17.37%	25.58%	16.91%	15.82%	27.96%	12.92%
SAT	3.97%	6.14%	4.32%	3.32%	6.54%	2.57%	3.37%	5.41%	2.25%

种黑盒攻击方法（One-pixel、NES(PI) 和 POBA-GA）直接对 6 个模型进行攻击。同样，表 4 中的每一种攻击方法下都对应应有被允许的最大模型访问次数。从实验结果可知，在面对基于概率优化的不同黑盒攻击方法时，SAT 相比其他防御方法具有更强的鲁棒性。具体实验结果如表 4 所示。

由表 4 可知，除 One-pixel 攻击方法外，其他两种防御方法在原始模型上的攻击成功率近似于 4.2 节中的 ZOO、Auto-ZOO 和 Boundary++ 攻击。但 One-pixel 的成功率远远低于其他两种方法，这是因为 One-pixel 攻击方法通过只改变若干个像素点（本节实验中小于 5 个）来生成对抗样本，对于 224×224 的 ImageNet 图像，扰动值过小。过小的扰动很容易通过预处理操作被除去，所以 Transform 在 One-pixel 上的防御效果远优于其他两种攻击方法。由于 RSE 在防御过程中均引入随机性扰动，因此在基于概率优化的黑盒攻击防御

性能明显优于除 SAT 外其余防御算法，但引入的随机扰动会对原始模型的分类准确率造成较大的负面影响。表 2 中 RES 在 MNIST 和 ImageNet 数据集中对 One-pixel 攻击的防御效果相比 SAT 更佳，进一步验证了 RES 防御带入的随机不确定性造成的不稳定性，而 SAT 则不会有此问题。最后由表 2~表 4 可知，SAT 在这 3 类黑盒攻击上的防御效果稳定优于其他防御方法，为本文方法的理论论证提供了实验支撑。

4.5 SAT 与其他防御方法的防御代价对比

本节实验主要分析 SAT 与其他防御方法的防御代价对比，主要比较防御后模型的原始识别准确率（ACC）、模型训练时间（ T_{train} ）和模型推理时间（ T_{test} ）。攻击方法使用基于梯度估计的黑盒攻击 ZOO，具体结果如表 5 所示。防御代价包括模型精度损失代价和防御时间代价。模型精度损失通过比较防御前后的原始模型识别准确率，以表 5 中的 ImageNet 数据集为例，SAT 方法的 ACC

表 4 基于概率优化的黑盒攻击防御效果对比
Table 4 Comparison of black box attack defense effects based on probability optimization

模型	MNIST (ASR)			CIFAR10 (ASR)			ImageNet (ASR)		
	One-pixel (80 000)	NES(PI) (2 000)	POBA-GA (1 000)	One-pixel (80 000)	NES(PI) (2 000)	POBA-GA (1 000)	One-pixel (800 000)	NES(PI) (50 000)	POBA-GA (5 000)
原始模型	80.32%	99.26%	100.00%	86.43%	100.00%	100.00%	40.61%	93.60%	98.00%
AdvTrain	30.12%	33.81%	41.86%	35.29%	38.30%	38.52%	9.43%	32.96%	38.10%
Transform	4.54%	25.48%	20.55%	8.19%	24.88%	19.96%	2.48%	28.90%	18.54%
Ensemble	13.65%	26.73%	27.43%	15.3%	23.65%	26.92%	8.54%	22.49%	23.95%
RSE	1.43%	15.54%	8.93%	7.69%	20.52%	16.49%	2.43%	18.97%	10.80%
SAT	3.69%	7.31%	6.32%	3.91%	6.11%	7.32%	4.56%	5.23%	6.43%

表 5 SAT 与其他防御方法的防御代价对比
Table 5 Comparison of defense costs between SAT and other defense methods

模型	MNIST			CIFAR10			ImageNet		
	ACC	$T_{\text{Train}}/\text{min}$	T_{Test}/s	ACC	$T_{\text{Train}}/\text{min}$	T_{Test}/s	ACC	$T_{\text{Train}}/\text{h}$	T_{Test}/s
原始模型	100.00%	10.34	24.54	90.49%	24.13	24.18	76.53%	26.45	42.46
AdvTrain	98.88%	24.35	25.36	82.51%	49.87	24.66	70.59%	50.53	43.52
Transform	99.96%	12.39	26.16	88.64%	25.21	25.31	72.18%	28.47	48.38
Ensemble	99.98%	30.24	41.54	92.87%	62.59	42.10	76.94%	61.87	130.21
RSE	98.83%	26.85	37.63	87.58%	57.37	39.51	66.48%	56.40	75.19
SAT	100.00%	10.34	25.18	90.49%	24.13	24.01	76.53%	26.45	42.79

没有下降,而 AdvTrain 等方法的 ACC 下降平均超过 6.5%,ACC 下降越多,说明代价越大。防御时间代价通过比较防御前后的训练时间和推理时间,以表 5 中的 ImageNet 数据集为例,SAT 防御后的训练时间和推理时间增长分别为 0 和 0.7%,而 AdvTrain 等方法的训练时间和推理时间平均增长了 86%和 55%,训练和推理时间越长,说明防御代价越大。

由表 5 可知,SAT 与原始模型的准确率和训练时间都相同,推理时间也相近,这是因为 SAT 不需要对模型进行额外的训练,也不会改变概率最高的类标从而影响原模型的准确率,换言之,SAT 因为防御效果而额外产生的防御代价为 0。防御效果仅次于 SAT 的 RSE,其训练时间远高于 SAT,且在 ImageNet 大数据集上的准确率下降明显。除以上两者之外,其他 3 种防御方法在防御效果上不如 SAT,而且防御代价明显高于 SAT。由上所述,SAT 是一种快速有效的轻量级防御方法。

4.6 SAT 抵御黑盒攻击的有效性分析

本节以 MNIST 和 CIFAR10 数据集为例,基于 t -SNE 图对 SAT 防御前后的样本分布概率进行可视化,并计算输出概率的类间距离。理论上,经过 SAT 进行重新激活后,样本的输出概率分布变得十分接近,直观的体现是在分布图上的样本变得更加混乱。但是 SAT 的单调性保证了正常样本的正确识别。进一步,本节使用类间距离对 SAT 防御前后的输出概率的类间距离进行计算,理论上类间距离会因为 SAT 的操作而变得接近,令黑盒攻击误以为能够以较小的代价很容易地实施攻击,但生成的对抗样本在目标模型中是无效的。

本节分别从 MNIST 和 CIFAR10 数据集选取了 2 000 张正常样本,输入对应的目标模型,分别保存 softmax 激活后输出的概率向量和 SAT 激活后输出的概率向量;然后使用 t -SNE 降维方法对概率向量进行降维并绘制二维可视图,如图 5 所示。可以直观地看出,经过 SAT 激活后,样本点的分布变得混乱了。同时,MNIST 数据集的 ResNet50 模型上,类间距离由 1.253 降低为 0.016;CIFAR10 数据集的 ResNet50 模型上,类间距离由

0.976 降低为 0.010。通过增加输出概率分布的混乱程度,减少了模型概率信息的泄露,实现了有效的对抗防御。

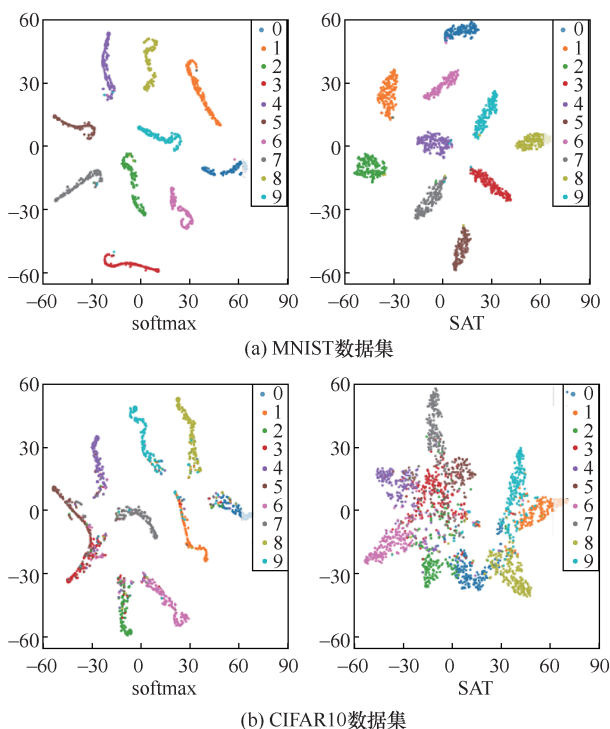


图 5 不同数据集的目标模型在 SAT 防御前后的 t -SNE 分布
Figure 5 t -SNE distribution of target models of different datasets before and after SAT defense

4.7 敏感性分析

本节主要分析动态隐私预算的变化区间对防御效果的影响。具体实验结果如图 6 所示,其中,横坐标表示可选的动态变化区间,纵坐标表示攻击成功率,每一个区间都展示了 4.2 节至 4.4 节中 9 种攻击方法的攻击。

由图 6 可知,基于概率优化的 3 种黑盒攻击方法的攻击成功率与区间的范围大小有关,与区间中值的大小关系不明显,区间的范围越大,攻击成功率越低,表示防御效果越好。基于梯度估计和模型等价的 6 种攻击方法的攻击成功率与区间中值的大小有关,与区间的范围大小关系不明显,当区间中值越小时,攻击成功率越低,表示防御效果越好。但总体上,SAT 的防御效果对区间的变化并没有非常敏感,为了同时兼顾 3 类黑盒攻击方法的防御效果,前文实验中 ε 选取的区间为 $[0,1]$ 。

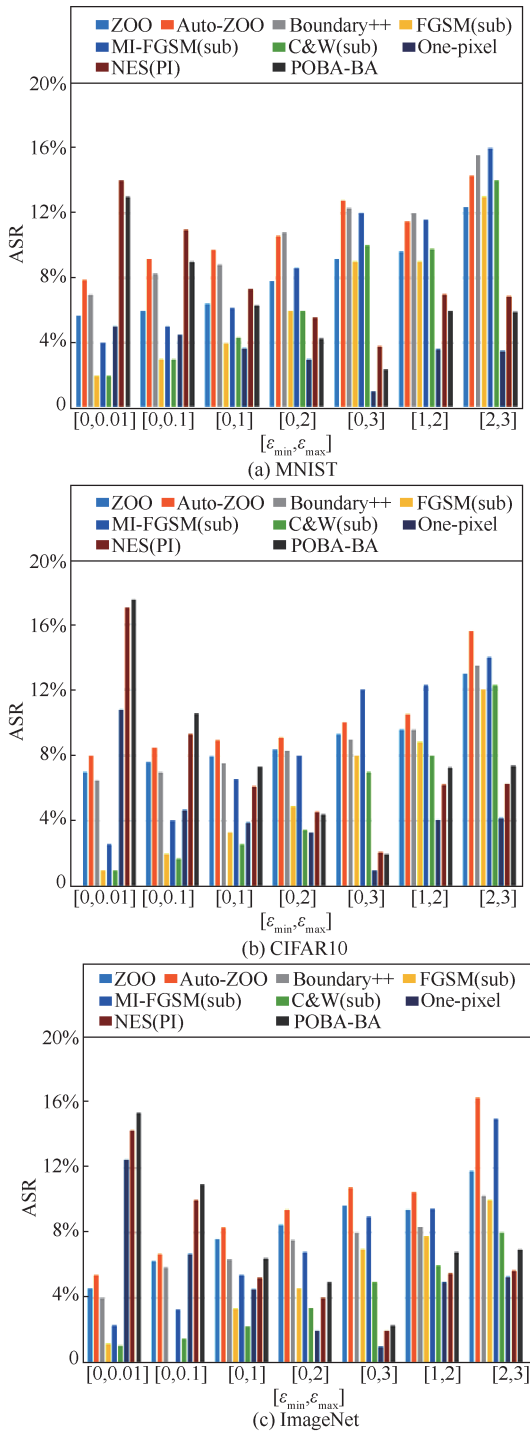


图 6 动态隐私预算的变化区间敏感性分析
Figure 6 Sensitivity analysis chart of the change interval of dynamic privacy budget

5 结束语

本文提出了一种基于 softmax 激活变换的攻击无关的可验证防御方法，该方法能够对黑盒攻击实现快速、轻量的防御。所提方法是一

种可验证的防御方法，能够基于概率信息隐私保护和 softmax 激活变换推导其防御的有效性和可靠性。SAT 不参与目标模型的训练，仅在推理阶段通过对目标模型的输出概率进行重新激活，实现概率信息的隐私保护从而防御黑盒攻击。SAT 与对抗攻击无关，避免了制作大量对抗样本的负担，实现了攻击的事前防御。SAT 的激活过程具有单调性，因此不会影响正常样本的识别。本文为 SAT 设计了可变隐私保护变换系数策略，实现了动态防御。实验结果表明，本文提出的对抗防御方法在降低攻击成功率、防御成本、参数敏感性方面优于许多已有的对抗防御方法，因此，本文方法是可行且高效的。

在实验过程中，发现 SAT 存在一定的问题，即在面对基于类标的黑盒攻击时，SAT 无法实现很好地防御，这是 SAT 不改变模型内部结构也不影响输出类的最大类标造成的。在未来的研究中，将进一步探索针对基于类标的黑盒攻击的高效防御方法。

参考文献：

- [1] GOODFELLOW I, BENGIO Y, COURVILLE A. Deep learning[M]. MIT Press, 2016.
- [2] HELMSTAEDTER M, BRIGGMAN K L, TURAGA S C, et al. Connectomic reconstruction of the inner plexiform layer in the mouse retina[J]. Nature, 2013, 500(7461): 168-174.
- [3] XIONG H Y, ALIPANAHI B, LEE L J, et al. The human splicing code reveals new insights into the genetic determinants of disease[J]. Science, 2015, 347(6218): 1254806-1254815.
- [4] HINTON G E, DENG L, YU D, et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups[J]. IEEE Signal Processing Magazine, 2012, 29(6): 82-97.
- [5] SUTSKEVER I, VINYALS O, LE Q V, et al. Sequence to sequence learning with neural networks[J]. Advances in neural information processing systems, 2014, 4(2): 3104-3112.
- [6] KRIZHEVSKY A, SUTSKEVER I, HINTON G E, et al. ImageNet classification with deep convolutional neural networks[J]. Advances in Neural Information Processing Systems, 2012, 2: 1097-1105.
- [7] DENG J, DONG W, SOCHER R, et al. ImageNet: a large-scale hierarchical image database[C]//IEEE Conference on Computer Vision and Pattern Recognition. 2009: 248-255.
- [8] 田野, 项世军. 基于 LBP 和多层 DCT 的人脸活体检测算法[J]. 计算机研究与发展, 2018, 55(3): 643-650.
- TIAN Y, XIANG S J. LBP and multilayer DCT based anti-spoofing countermeasure in face liveness detection[J]. Journal of Computer

- Research and Development, 2018, 55(3): 643-650.
- [9] 张蕊, 李锦涛. 基于深度学习的场景分割算法研究综述[J]. 计算机研究与发展, 2020, 57(4): 859-875.
ZHANG R, LI J T. A survey on algorithm research of scene parsing based on deep learning[J]. Journal of Computer Research and Development, 2020, 57(4): 859-875.
- [10] NAJAFABADI M M, VILLANUSTRE F, KHOSHGOFTAAR T M, et al. Deep learning applications and challenges in big data analytics[J]. Journal of Big Data, 2015, 2(1): 1-21.
- [11] PAPERNOT N, MCDANIEL P, SINHA A, et al. Towards the science of security and privacy in machine learning[J]. arXiv preprint arXiv:1611.03814, 2016.
- [12] MNIH V, KAVUKCUOGLU K, SILVER D, et al. Human-level control through deep reinforcement learning[J]. Nature, 2015, 518(7540): 529-533.
- [13] SZEGEDY C, ZAREMBA W, SUTSKEVER I, et al. Intriguing properties of neural networks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR). 2013: 7185-7193.
- [14] GOODFELLOW I J, SHLENS J, SZEGEDY C. Explaining and harnessing adversarial examples[J]. arXiv preprint arXiv:1412.6572, 2014.
- [15] DONG Y, LIAO F, PANG T, et al. Boosting adversarial attacks with momentum[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR). 2018: 9185-9193.
- [16] MOOSAVI-DEZFOOLI S M, FAWZI A, FAWZI O, et al. Universal adversarial perturbations[C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2017: 1765-1773.
- [17] CARLINI N, WAGNER D. Towards evaluating the robustness of neural networks[C]//Proceedings of 2017 IEEE European Symposium on Security and Privacy. 2017: 39-57.
- [18] CHEN J, ZHENG H, XIONG H, et al. MAG-GAN: massive attack generator via GAN[J]. Information Sciences, 2020, 536: 67-90.
- [19] SU J, VARGAS D V, SAKURAI K. One pixel attack for fooling deep neural networks[J]. IEEE Transactions on Evolutionary Computation, 2019, 23(5): 828-841.
- [20] CHEN P Y, ZHANG H, SHARMA Y, et al. Zoo: zeroth order optimization based black-box attacks to deep neural networks without training substitute models[C]//Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security. 2017: 15-26.
- [21] CHEN J, SU M, SHEN S, et al. POBA-GA: perturbation optimized black-box adversarial attacks via genetic algorithm[J]. Computers & Security, 2019, 85: 89-106.
- [22] 陈晋音, 陈治清, 郑海斌, 等. 基于粒子群优化的路牌识别模型的黑盒物理攻击方法[J]. 软件学报, 2020: 1-17.
CHEN J Y, CHEN Z Q, ZHENG H B, et al. Black-box physical attack against road sign recognition model via PSO[J]. Journal of Software, 2020: 1-17.
- [23] 陈晋音, 沈诗婧, 苏蒙蒙, 等. 车牌识别系统的黑盒对抗攻击[J]. 自动化学报, 2020: 1-18.
CHEN J Y, SHEN S J, SU M M, et al. Black-box adversarial attack on license plate recognition system[J]. Acta Automatica Sinica, 2020: 1-18.
- [24] BRENDEN W, RAUBER J, BETHGE M. Decision-based adversarial attacks: reliable attacks against black-box machine learning models[J]. arXiv preprint arXiv:1712.04248, 2017.
- [25] CHEN J, ZHENG H, CHEN R, et al. RCA-SOC: a novel adversarial defense by refocusing on critical areas and strengthening object contours[J]. Computers & Security, 2020: 101916.
- [26] AKHTAR N, MIAN A. Threat of adversarial attacks on deep learning in computer vision: a survey[J]. IEEE Access, 2018, 6: 14410-14430.
- [27] ALENAZY W M, ALQAHTANI A S. Gravitational search algorithm based optimized deep learning model with diverse set of features for facial expression recognition[J]. Journal of Ambient Intelligence and Humanized Computing, 2021, 12(2): 1631-1646.
- [28] TU C, TING P, CHEN P Y, et al. Autozoom: Autoencoder-based zeroth order optimization method for attacking black-box neural networks[J]. arXiv preprint arXiv:1805.11770, 2018.
- [29] CHEN J, JORDAN M I. Boundary attack++: query-efficient decision-based adversarial attack[J]. arXiv preprint arXiv:1904.02144, 2019.
- [30] PAPERNOT N, MCDANIEL P, GOODFELLOW I, et al. Practical black-box attacks against machine learning[C]//Proceedings of the 2017 ACM on Asia Conference on Computer and Communications. 2017: 506-519.
- [31] ZHANG X, WU D. On the vulnerability of CNN classifiers in EEG-based BCIs[J]. IEEE Transactions on Neural Systems and Rehabilitation Engineering, 2019, 27(5): 814-825.
- [32] SU J, VARGAS D V, SAKURAI K. One pixel attack for fooling deep neural networks[J]. IEEE Transactions on Evolutionary Computation, 2019, 23(5): 828-841.
- [33] ILYAS A, ENGSTROM L, ATHALYE A, et al. Black-box adversarial attacks with limited queries and information[C]//Proceedings of 35th International Conference on Machine Learning(ICML). 2018: 3392-3401.
- [34] AKHTAR N, MIAN A. Threat of adversarial attacks on deep learning in computer vision: a survey[J]. IEEE Access, 2018, 6: 14410-14430.
- [35] GOODFELLOW I J, SHLENS J, SZEGEDY C. Explaining and harnessing adversarial examples[J]. arXiv preprint arXiv:1412.6572.
- [36] KURAKIN A, GOODFELLOW I, BENGIO S, et al. Adversarial machine learning at scale[C]//Proceedings of Computer Vision and Pattern Recognition(CVPR). 2016: 1-13.
- [37] MADRY A, MAKELOV A, SCHMIDT L, et al. Towards deep learning models resistant to adversarial attacks[C]//ICLR. 2018.
- [38] MADRY A, MAKELOV A, SCHMIDT L, et al. Towards deep learning models resistant to adversarial attacks[J]. arXiv preprint arXiv:1706.06083, 2017.
- [39] SHAFABI A, NAJIBI M, GHASI A, et al. Adversarial training for free![J]. arXiv preprint arXiv: 1904.12843v1.
- [40] XIE C, WANG J, ZHANG Z, et al. Adversarial examples for semantic segmentation and object detection[C]//Proc of International Conference on Computer Vision(ICCV). 2017: 1-13.
- [41] GUO C, RANA M, CISSE M, et al. Countering adversarial images using input transformations[J]. arXiv preprint arXiv:1711.00117, 2017.

- [42] STRAUSS T, HANSELMANN M, JUNGINGER A, et al. Ensemble methods as a defense to adversarial perturbations against deep neural networks[J]. arXiv preprint arXiv:1709.03423, 2017.
- [43] LIU X, CHENG M, ZHANG H, et al. Towards robust neural networks via random self-ensemble[C]//Proceedings of Lecture Notes in Computer Science, 2018, 11211: 381-397.



吴长安（1996- ），男，浙江长兴人，浙江工业大学硕士生，主要研究方向为深度学习、计算机视觉和图像的对抗攻击和防御。

[作者简介]



陈晋音（1982- ），女，浙江象山人，浙江工业大学教授，主要研究方向为人工智能安全、图数据挖掘和进化计算。



郑海斌（1995- ），男，浙江台州人，浙江工业大学博士生，主要研究方向为深度学习、人工智能安全和图像识别。