软件学报 ISSN 1000-9825, CODEN RUXUEW
2025,36(10):4864−4879 [doi: 10.13328/j.cnki.jos.007300] [CSTR: 32375.14.jos.007300]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

基于图像变换的双阈值对抗样本检测*

刘 会[1,2], 文福举[1,2], 杜红琴[1,2], 王敬华[1,2], 赵 波[3]

[1](华中师范大学 计算机学院, 湖北 武汉 430079)

[2](人工智能与智慧学习湖北省重点实验室 (华中师范大学), 湖北 武汉 430079)

[3](武汉大学 国家网络安全学院, 湖北 武汉 430072)

通信作者: 赵波, E-mail: zhaobo@whu.edu.cn

摘 要: 当前基于图像变换的对抗样本检测方法利用了图像变换对对抗样本的特征分布造成较大的影响, 而对于良性样本的特征分布影响较小这一特点, 通过计算样本变换前后的特征距离来检测对抗样本. 然而随着对抗攻击的深入研究, 研究者们更注重加强对抗攻击的鲁棒性, 使得一些攻击能“免疫”图像变换带来的影响. 现有方法难以有效地检测出鲁棒性强的对抗样本. 发现当前的对抗样本过于鲁棒, 强鲁棒性对抗样本在图像变换下的特征分布距离远小于良性样本的特征分布距离, 其特征分布距离违背了良性样本特征分布规律. 基于这一关键的发现, 提出基于图像变换的双阈值对抗样本检测方法, 在传统单阈值检测方法的基础上设置一个下阈值, 构成双阈值检测区间, 其特征分布距离不在区间范围的样本将被判定为对抗样本. 在 VGG19、DenseNet 和 ConvNeXt 图像分类模型中开展广泛的验证. 实验证明该方法能够有效兼容现有单阈值检测方案的检测能力, 同时对强鲁棒性对抗样本表现出良好的检测效果.

关键词: 图像变换; 对抗样本; 特征分布; 双阈值检测; 图像分类

中图法分类号: TP309

中文引用格式: 刘会, 文福举, 杜红琴, 王敬华, 赵波. 基于图像变换的双阈值对抗样本检测. 软件学报, 2025, 36(10): 4864–4879. http://www.jos.org.cn/1000-9825/7300.htm

英文引用格式: Liu H, Wen FJ, Du HQ, Wang JH, Zhao B. Dual-threshold Adversarial Example Detection Based on Image Transformation. Ruan Jian Xue Bao/Journal of Software, 2025, 36(10): 4864–4879 (in Chinese). http://www.jos.org.cn/1000-9825/7300.htm

Dual-threshold Adversarial Example Detection Based on Image Transformation

LIU Hui[1,2], WEN Fu-Ju[1,2], DU Hong-Qin[1,2], WANG Jing-Hua[1,2], ZHAO Bo[3]

[1](School of Computer Science, Central China Normal University, Wuhan 430079, China)

[2](Hubei Provincial Key Laboratory of Artificial Intelligence and Smart Learning (Central China Normal University), Wuhan 430079, China)

[3](School of Cyber Science and Engineering, Wuhan University, Wuhan 430072, China)

Abstract: Existing adversarial example detection methods based on image transformation employ the characteristic that the image transformation can significantly change the feature distribution of adversarial examples but slightly change the feature distribution of benign examples. Adversarial examples can be detected by calculating the feature distance before and after image transformation. However, with the deepening research on adversarial attacks, researchers pay more attention to enhancing the robustness of adversarial examples, so that some attacks can be “immune” to the effect exerted by image transformation. Existing methods are difficult to detect robust adversarial examples effectively. This paper observes that the existing adversarial examples are too robust, and the feature distribution distance of robust adversarial examples under image transformation is much smaller than that of benign examples, which is not consistent

* 基金项目: 国家资助博士后研究人员计划 (GZC20230922); 中国博士后科学基金第 75 批面上项目 (2024M751050); 华中师范大学中央高校基本科研业务费 (CCNU24XJ001); 华中师范大学基本科研业务费 (自然科学类)(CCNU24ai010)
收稿时间: 2024-05-06; 修改时间: 2024-07-14, 2024-08-29, 2024-09-19; 采用时间: 2024-10-02; jos 在线出版时间: 2025-01-24
CNKI 网络首发时间: 2025-01-26

with the feature distribution laws of benign examples. Based on this key observation, this study proposes a dual-threshold adversarial example detection based on image transformation, which sets a lower threshold combining existing single-threshold methods to form a dual-threshold detection interval. An example whose feature distribution is not within the dual-threshold detection interval will be judged as an adversarial example. Additionally, this study conducts extensive experiments on VGG19, DenseNet, and ConvNeXt models for image classification. The results show that the proposed approach is compatible with the detection ability of existing single-threshold detection schemes, and yields outstanding detection performance against robust adversarial examples.

Key words: image transformation; adversarial example; feature distribution; dual-threshold detection; image classification

近年来, 深度学习发展迅速, 凭借其优异的性能被广泛应用于计算机视觉[1]、语音识别[2]、自然语言处理[3]、推荐系统[4]等领域. 通过对大规模数据的复杂特征进行抽取学习, 深度学习模型能够针对这些领域的目标任务提供更准确高效的个性化解决方案, 推动这些领域的发展和创新. 尽管如此, 深度学习应用中的安全性问题依旧不容忽视. 早在 Szegedy 等人[5]的研究中, 就发现了对抗性样本带来的安全问题. 对抗样本以微小扰动误导深度学习模型做出错误决策, 对深度学习模型在自动驾驶[6]、生物识别[7]、医疗成像[8]等关键领域的应用安全构成了严峻的威胁.

为了保障深度学习技术在实际应用中的安全性, 研究者提出了许多对抗样本防御措施, 如对抗训练[9]、梯度遮掩[10]、数据预处理[11]等. 对抗训练通过纳入对抗样本至模型训练过程中, 有效增强了模型鲁棒性, 但是需要大量对抗样本数据, 容易产生过拟合. 梯度遮掩则通过隐藏模型梯度信息增加生成对抗样本的难度, 但难以防御不依赖于梯度信息的攻击方法. 数据预处理通过对输入样本进行降噪处理以消除对抗噪声, 然而有时却因此损失了重要特征信息, 造成模型性能受损. 鉴于以上防御措施的种种限制, 研究人员将防御对抗样本的注意力转向了对抗样本检测[12]方向. 对抗样本检测旨在通过对输入数据进行检测, 识别出对抗样本, 进而有效地抵御对抗攻击. 该防御方式通常不需要更改模型或者样本的信息, 即可保证深度学习技术在应用过程中的数据完整性.

图像变换是对抗样本检测的重要方法之一. 通过分析样本在变换前后的特征分布差异, 并设定阈值判断标准, 可以有效区分良性样本与对抗样本. 这些方案利用了对抗样本较良性样本对输入变换相对敏感的特点. 然而, 为了提升对抗样本的逃逸能力, 基于对模型漏洞和脆弱性的深刻理解, 精准地设计出在变换后仍能高置信度欺骗目标模型的对抗样本, 显著提升了对抗样本的鲁棒性. 这些鲁棒性较高的对抗样本有潜力规避上述检测方法, 有效欺骗目标深度学习模型. 然而我们发现, 精心设计的高鲁棒性对抗样本对于图像变换表现的过于稳定, 其变换前后的特征分布距离远小于良性样本的特征分布距离. 基于这一关键的观察, 我们设置了下阈值来检测鲁棒性高的对抗样本, 结合传统的单阈值检测方法, 有效提升对于高鲁棒性对抗样本的检测能力.

本文提出了一种基于图像变换的双阈值对抗样本检测方法. 该方法首先测算样本在图像变换前后的特征分布距离. 在传统的单阈值基础上, 引入下阈值构建出阈值区间, 评判样本在变换前后特征分布距离是否落于该区间内, 进而确定其为良性或对抗样本. 实验结果表明, 本文方法对鲁棒性较强和鲁棒性较弱的对抗样本, 均表现出良好的检测效果. 本文的主要贡献如下.

(1) 我们观察到当前先进的对抗样本攻击为逃逸检测, 生成的对抗样本鲁棒性较高, 对图像变换表现得过于稳定, 往往违背了良性样本特征分布的统计规律, 从而为高鲁棒性对抗样本的有效检测提供了新的思路.

(2) 本文提出了基于图像变换的双阈值对抗样本检测框架, 通过在传统的单阈值对抗样本检测方法的基础上, 设置下阈值来检测鲁棒性较强的对抗样本, 显著提升了当前单阈值检测器对高鲁棒性对抗样本的检测能力.

(3) 实验评估了本文所提的双阈值检测方法面向 VGG19、DenseNet 和 ConvNeXt 图像分类模型, 基于 9 种图像变换对 5 种基准对抗样本攻击的检测性能, 并与其他先进的检测方案进行对比. 实验证明, 本文所提方法能够有效弥补当前单阈值检测的局限性.

本文第 1 节介绍关于对抗样本检测领域的相关工作. 第 2 节介绍对抗攻击以及分类模型的背景知识. 第 3 节重点介绍本文的研究动机. 第 4 节介绍双阈值对抗样本检测器的设计方法. 第 5 节介绍实验设置以及对实验结果进行分析. 第 6 节对全文做出总结.

1 相关工作

近年来, 已有大量针对对抗样本检测领域的相关研究, 其检测方法大致上可分为以下 4 类: 基于统计方法的对抗样本检测、基于辅助模型的对抗样本检测、基于神经网络特性的对抗样本检测和基于输入变换的对抗样本检测.

基于统计方法的对抗样本检测通过分析良性样本与对抗样本的统计特征, 找出二者之间的统计差异, 从而识别对抗样本. 如 Hendrycks 等人[13]发现, 分类模型在推断对抗样本与良性样本时, 其输出的置信度分布存在较大的差异. 良性样本的置信度分布远离均匀分布, 而对抗样本的置信度分布靠近均匀分布, 因此可以通过计算其与均匀分布的 KL 散度来检测对抗样本. Feinman 等人[14] 则认为良性样本与对抗样本数据分布存在差异, 利用训练样本的 logits 分布与对应的预测标签来确定样本的数据流形, 通过高斯核密度估计来判断待测样本与目标类别流形的距离, 核密度估计值越小, 表明与目标类别流形的距离越远.

基于辅助模型的对抗样本检测主要通过学习良性样本与对抗样本之间的特征差异来进行检测. 按训练数据是否需要手动标注, 检测方法可分为有监督与无监督两类. 有监督的辅助模型需要同时学习良性样本与对抗样本的特征. 如 Gong 等人[9]利用良性样本与对抗样本来训练一个二元分类器, 使其能有效区分对抗样本. Lust 等人[15]以梯度范数作为指标进行对抗样本检测. 该方法检测开销小、速度快, 但检测过程依赖梯度信息, 对于非梯度攻击的对抗样本检测效果差. 基于无监督的辅助模型仅利用良性样本的特征构建检测模型. 如 Liu 等人[16]利用良性样本的二元特征集训练孤立森林, 从而实现无监督的对抗样本检测. Wang 等人[17] 提出 ADDITION 模型, 通过注入与图像相关噪声, 将对抗扰动转化为近似高斯噪声, 再使用降噪处理消除对抗扰动, 最后通过分类不一致性来识别对抗样本.

基于神经网络特性的对抗样本检测通过观察良性样本与对抗样本在神经网络中的表现差异进行检测, 主要利用神经元的激活值或者输出值. Zheng 等人[18]提出的 I-defender 方法, 通过建模良性样本在隐层神经元的输出分布来检测对抗样本. 该方法检测能力强, 但计算开销大、效率低. Eniser 等人[19]提出了 RAID 检测方法, 利用良性样本与对抗样本的神经元激活值差异训练二分类检测器, 并通过池化的 RAID 增强检测器的鲁棒性. Ma 等人[20]通过观察深度神经网络中激活值分布变化, 利用起源不变量和激活值不变量模型来检测对抗样本.

基于输入变换的对抗样本检测方法利用对抗样本较良性样本对输入变换相对敏感的特点, 通过对输入数据进行各种变换操作, 比较目标模型对变换前后样本的输出结果, 来识别对抗样本. Tian 等人[21]发现对抗样本对平移和旋转很敏感, 他们通过对输入样本进行不同程度的平移和旋转操作获取其分类模型输出的 logits 分布, 训练检测器检测对抗样本, 实验表明他们的方法对 CW 攻击有着很好的检测效果. Liang 等人[11]提出了自适应降噪方法, 使用标量量化和平滑滤波技术, 通过比较降噪前后的模型预测不一致性来检测是否为对抗样本. 该方法无需参考任何攻击先验知识但是低置信度良性样本易被误判. Ryu 等人[22]则提出基于图像熵的对抗检测方法, 通过对输入样本进行位深度减少处理, 计算样本图像熵变化, 若变化过大则为对抗样本.

当前对抗样本检测领域已经取得了一定进展. 然而, 随着对抗样本攻击技术的不断更新, 对抗样本的鲁棒性显著增强, 对抗样本的逃逸能力得到了明显的提升, 现有的检测技术难以准确识别强鲁棒性的对抗样本. 本文针对强鲁棒性对抗样本检测困难问题, 提出了一种基于图像变换的双阈值检测方案, 克服了现有的单阈值检测方案在强鲁棒性对抗样本检测中的局限性, 显著提升了对抗样本检测效果.

2 背景知识

2.1 对抗攻击

对抗攻击是指通过对输入样本进行细微的、有针对性的扰动, 达到欺骗深度学习模型的行为. 这类扰动通常不会改变人眼观察的图像本质, 但可以使深度学习模型产生错误的输出结果. 下面介绍几种常见的对抗攻击方法.

(1) 快速梯度符号法 (the fast gradient sign method, FGSM). FGSM 攻击最开始由 Goodfellow 等人[23]提出, 是一

种基于梯度的单次攻击方法. 它利用目标模型的梯度信息生成对抗扰动并添加到输入样本中生成对抗样本. 其攻击的形式化表达如公式 (1) 所示:

$$x_{\text{adv}} = x + \varepsilon \cdot Sign(\nabla_x J(\theta, x, y)) \tag{1}$$

其中, x_{adv} 是对抗样本, x 是输入样本, y 是 x 对应的标签, ε 是扰动大小, θ 是模型的参数, $Sign(\cdot)$ 表示取梯度的符号, $J(\theta, x, y)$ 是关于模型的损失函数, $\nabla_x J(\theta, x, y)$ 是计算样本相对于损失函数的梯度.

(2) 基本迭代方法 (basic iterative method, BIM). BIM 攻击[24]是一种常见的迭代对抗攻击方法, 它是对 FGSM 攻击的改进与扩展. BIM 通过在每次迭代过程中添加由梯度决定的小扰动, 并在原始样本的一定范围内进行裁剪修改, 逐步逼近目标攻击效果. 该方法可以更好地探索目标模型的敏感区域, 生成扰动更小的对抗样本, BIM 攻击的形式化表达如公式 (2) 所示:

$$x_{\text{adv}}^{N+1} = Clip_{\varepsilon}\left\{x_{\text{adv}}^{N} + \alpha \cdot Sign(\nabla_x J(\theta, x_{\text{adv}}^{N}, y))\right\} \tag{2}$$

其中, x_{adv}^{N} 表示第 N 次迭代的版本, α 表示每次移动的步长, $Clip_{\varepsilon}$ 表示在扰动范围 ε 内进行裁剪.

(3) 投影梯度下降攻击 (projected-gradient descent attack, PGD). PGD 攻击[25]是 FGSM 攻击的多步变体, 它通过采用投影梯度下降的方式处理对抗样本生成的优化问题. 在每次迭代的过程中 PGD 会沿着梯度的反方向按照设定的步长逐步更新样本, 然后将结果映射至原始扰动空间, 以确保其扰动大小不超过预设限制. PGD 攻击的形式化表达如公式 (3) 所示:

$$x_{\text{adv}}^{N+1} = \prod\nolimits_{x+s}\left\{x_{\text{adv}}^{N} + \alpha \cdot Sign(\nabla_x J(\theta, x_{\text{adv}}^{N}, y))\right\} \tag{3}$$

其中, s 是随机扰动, $\prod_{x+s}$ 是投影函数将值投影到 $x+s$ 的邻域范围内.

(4) DeepFool 攻击. DeepFool 攻击[26]不同于以上的 FGSM、PGD 等基于梯度的攻击, 其攻击原理基于解析几何. 在多分类问题中, DeepFool 认为分类边界和样本的距离即为改变分类标签的最小扰动, 通过不断迭代修改扰动, 可以将原始样本推向决策边界, 直至跨越分类边界实现攻击. 在相同的攻击成功率下, DeepFool 攻击生成的扰动比 FGSM 攻击的扰动更小, 攻击更为隐蔽.

(5) CW (Carlini and Wagner) 攻击. CW 攻击[27]是一种基于优化的攻击方法. 它遵循对抗攻击追求的两个关键目标, 即对抗样本与原始样本差距尽可能小和对抗样本能使模型以高置信度分类出错. CW 攻击将以上两个目标进行数学建模, 确定为两个目标函数, 通过最小化目标函数来寻找最优的攻击扰动. 该方法可以通过调节参数, 来控制生成对抗样本的扰动强度和置信度大小, 以此破解大部分对抗防御手段. 由于需要多轮迭代, CW 攻击生成对抗样本过程较为缓慢. CW 的 L_2 范数攻击可形式化表达为公式 (4):

$$\begin{cases} \min \|\delta\|_2 + c \cdot f(x_{\text{adv}}) \\ \text{s.t. } x_{\text{adv}} = x + \delta \in D \end{cases} \tag{4}$$

其中, δ 是扰动噪声, c 是一个超参数用来权衡两个损失函数之间的关系, $f(x_{\text{adv}})$ 定义见公式 (5):

$$f(x_{\text{adv}}) = \max(\max\{Z(x_{\text{adv}})_i : i \neq t\} - Z(x_{\text{adv}})_t, -k) \tag{5}$$

其中, $Z(\cdot)$ 是模型的 Softmax 层输出, t 是目标类别, k 是控制置信度的超参数.

2.2 图像分类模型

深度神经网络依然成为图像分类的主流方法. 基于深度神经网络的图像分类模型可以通过端到端的训练, 自主从输入图像数据中学习图像的特征表示和分类决策. 其中卷积神经网络 (convolutional neural network, CNN) 是最常用于图像分类任务的深度学习模型, 它通过多层卷积和池化来提取图像数据中的局部特征, 然后通过全连接层进行分类决策. 相比于传统的图像分类模型, CNN 能够自主学习到图像更深层次、更高级的特征表示, 从而做出更准确的图像分类决策. 在本文中, 我们在 VGG19[28]、DenseNet[29]和 ConvNeXt[30]这 3 个主流的 CNN 分类模型中开展验证试验.

VGG19 属于 VGGNet 系列, 由 Simonyan 等人[28]于 2014 年提出. 该模型凭借其深层结构和简明的网络架构, 在图像分类任务中展现出显著的优势. VGG19 由 16 个卷积层和 3 个全连接层组成, 19 个网络层分布在 6 个模块

中. 所有的卷积层中使用相同尺度的 3×3 的卷积核以及固定的步幅, 并在每个卷积层之后采用 2×2 的池化核进行下采样. 前 2 个模块中各含 2 个卷积层, 接下来的 3 个模块中各含 4 个卷积层, 而最后 1 个模块则由 3 个全连接层组成. 其中, 前 2 个全连接层包含 4 096 个神经元, 最终的全连接层由 1 000 个神经元组成. VGG19 通过使用多个小卷积核而非大卷积核来增加网络的深度, 以便捕捉更为复杂和高级的特征表示. 然而, 这种设计也带来了模型参数数量的显著增加以及更高的计算代价. 因此, 在模型的训练和推理过程中, VGG19 计算资源需求较大.

DenseNet 模型的核心组件是密集块 (dense block). 每个密集块由多个卷积层组成. 在密集块中, 每层均与之前所有层相连, 形成密集连接. 密集块内部所有的特征图都保持统一的维度, 不做尺寸裁剪. 另一个重要的组件是过渡层 (transition layer), 由 1×1 的卷积层和 2×2 的平均池化层构成, 用于减少特征图的尺寸和通道数, 从而限制模型的参数降低计算复杂度. DenseNet 的网络结构由多个密集块和过渡层组成, 最终通过一个 7×7 的平均池化层和全连接层进行分类决策. DenseNet 通过增强特征传播和改进梯度反向传播过程来减少参数数量. 与常规 CNN 不同, DenseNet 中每一层的输出不仅作为下一层的输入, 还直接与后续所有层的特征图相连. 这种密集连接结构有效提升了信息和梯度的传递, 增强了训练稳定性与特征利用率. 然而, 由于每层需保存所有前层的特征图, DenseNet 的内存消耗相对较大.

ConvNeXt 模型由 Liu 等人[30]于 2022 年提出, 是现代卷积神经网络的进化版本, 在图像分类任务中表现优异. 该模型结合了传统卷积操作与现代设计理念, 具有简洁的网络结构和高效的特征提取能力. ConvNeXt 由多个卷积块和归一化层组成, 采用标准的 3×3 卷积核和固定步幅, 并引入残差连接增强训练稳定性. 其前几个模块通过残差连接实现高效特征传递, 最后通过全局池化层和全连接层进行决策. 通过现代卷积操作和归一化机制, ConvNeXt 增加了网络深度和特征表示能力, 改善了梯度反向传播过程, 提高了模型训练效率. 尽管其设计导致了模型在内存和计算资源方面需求较大, 但其在图像分类、目标检测等任务上表现尤为突出.

3 研究动机

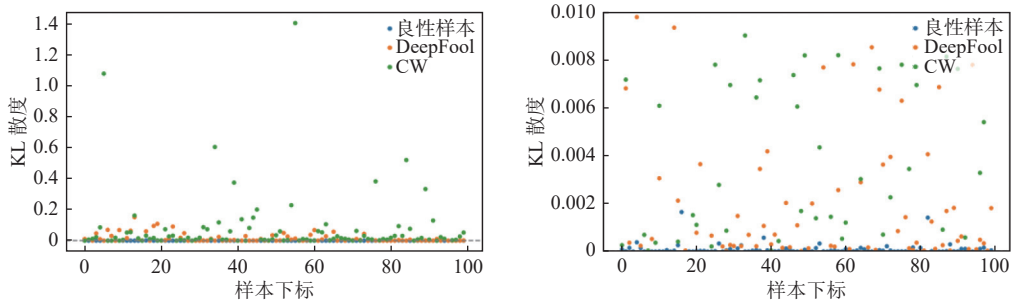
当前基于图像变换的对抗样本检测基于一个重要的观察: 面向图像变换, 神经网络分类模型对良性样本的推断结果较为稳定, 而对对抗样本的推断结果变化较大. 因此, 通过计算图像变换前后神经网络分类器输出结果的预测距离, 并设立一个合适的阈值, 可以有效检测对抗样本. 例如, 特征压缩技术[31]通过减少输入数据的位深度或应用空间平滑滤波, 旨在减轻对抗性干扰. 随后, 通过比较分类模型对原始输入与处理后输入预测结果的差异与预设阈值, 识别出超出阈值的样本作为对抗样本. 基于图像重构的检测策略[16]通过训练自编码器对输入图像数据进行编解码处理, 然后计算重构前后输入至分类模型的预测响应差异, 并与设定阈值比较, 以此判断样本是否为对抗样本.

大多数对抗样本在经过上述图像变换后, 神经网络分类器的预测结果会出现明显的变化, 产生较大的预测距离, 从而被当前检测方法有效识别. 为绕过这些检测手段, 攻击者可能专门针对上述防御技术, 构造对图像变换表现稳定的强鲁棒性对抗样本. 即使在经历多种图像变换后, 这些对抗样本仍能以高置信度误导分类模型. 这些强鲁棒性对抗样本在图像变换前后的预测距离通常明显低于检测阈值, 因而能轻易绕过检测机制.

然而, 这类强鲁棒性对抗样本并非无懈可击, 它们对图像变化表现得过分稳定成为检测的突破点. 我们观察发现, 良性样本在经过图像处理后其特征会发生相对轻微的改变, 而经精心设计的强鲁棒性对抗样本, 由于对图像变换过于稳定, 造成其预测距离过小. 为了验证此假设, 我们以 KL 散度作为预测概率分布差异的指标, 量化良性样本与对抗样本之间预测概率分布差异. 选取 100 张良性样本及通过 PGD、BIM、CW 和 DeepFool 攻击生成的对抗样本进行研究. 样本经位深度减小处理即每个通道位深度由 8 位降至 7 位后, 我们计算了样本变换前后预测概率分布的差异值.

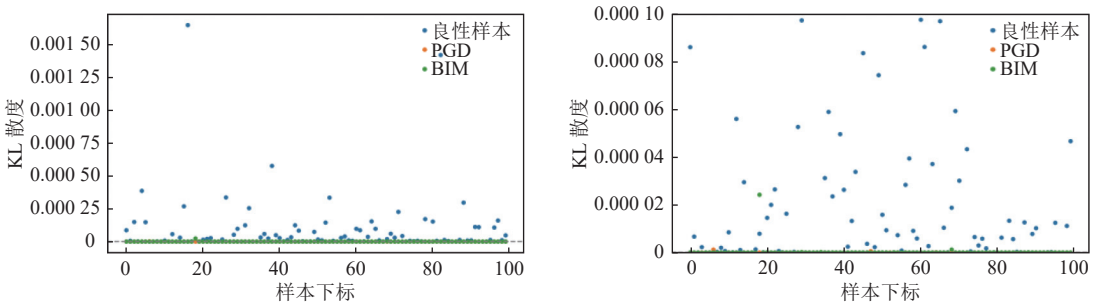
为了直观地观察对抗样本与良性样本之间的分布差异, 我们绘制了 KL 散度分布图. 由图 1 可以观察到良性样本与弱鲁棒性对抗样本的差异, 图 1(a) 展示的是 KL 散度在 0–1.4 之间的分布, 图 1(b) 特别显示了 KL 散度在 0–0.01 之间的样本分布情况. 对比图 1(a) 和图 1(b) 可以看到, 良性样本的 KL 散度分布主要集中在底部, 而 CW

和 DeepFool 对抗样本分布明显高于良性样本. 这一现象表明, 通过设置上阈值有助于检测鲁棒性较弱的对抗样本. 图 2 则展示了强鲁棒性对抗样本与良性样本之间的 KL 分布情况. 图 2 中对抗样本与良性样本之间 KL 散度分布差异较为明显, 其中 BIM 与 PGD 对抗样本的 KL 散度基本聚集在最底部, 良性样本分布则相对分散且位于两种对抗样本上层. 这一现象表明, 通过设置下阈值可以帮助我们检测鲁棒性较强的对抗样本. 基于以上观察, 我们认为通过设定上下双阈值来检测对抗样本可实现更全面的鉴别. 其中, 上阈值用以筛选鲁棒性较弱的对抗样本, 而下阈值则专门用于识别鲁棒性较强的对抗样本.



(a) KL 散度的全局分布情况 (b) KL ∈ [0, 0.01]的局部分布情况

图 1 弱鲁棒性对抗样本 KL 分布图



(a) KL 散度的全局分布情况 (b) KL ∈ [0, 0.0001]的局部分布情况

图 2 强鲁棒性对抗样本 KL 分布图

4 检测框架

从第 3 节的研究动机中我们发现, 随着对抗样本攻击技术的不断进步, 对抗样本的鲁棒性得到了显著的提升, 强鲁棒性对抗样本在面向图像变换时表现得异常稳定, 其预测距离明显低于良性样本. 因此, 本文在现有的单阈值对抗样本检测基础上, 提出用于检测强鲁棒性对抗样本的下阈值检测方法, 提出基于图像变换的双阈值对抗样本检测方法. 鲁棒性较弱的对抗样本经图像变换后, 分类模型输出的概率分布会出现显著差异, 其预测距离通常大于上阈值; 鲁棒性较强的对抗样本, 则表现为图像变换前后预测概率分布的高度一致性或仅有微小变化, 其预测距离通常小于下阈值. 而图像分类器对变换前后良性样本的预测距离通常处于特定的区间内. 相较于单阈值检测方法仅检测“弱鲁棒性”的对抗样本, 本文方法同时考虑“弱鲁棒性”和“强鲁棒性”两种特性. 该方法弥补了单阈值方法无法检测强鲁棒性对抗样本的局限性, 在检测能力上取得更优表现.

本方案是在传统的单阈值检测框架下新增一个下阈值, 构成双阈值区间来检测对抗样本, 其检测框架图如图 3 所示. 首先, 检测器对原始样本 x 进行图像变换得到变换后样本 x', 然后将 x 与 x' 输入分类模型进行预测得到对应的预测概率分布数组 P 和 Q, 计算 P 和 Q 之间的距离 d, 判断其是否在阈值区间 (τ_1, τ_2) 内. 若在该区间则判定输入样本为良性样本, 否则为对抗样本.

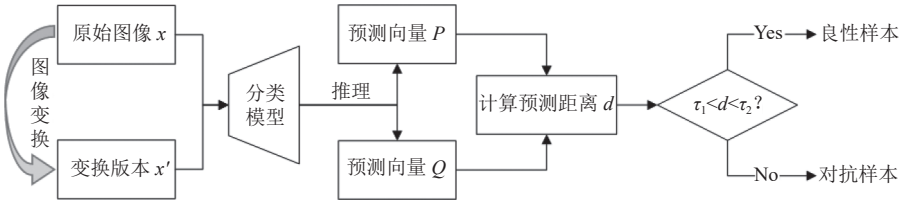


图 3　基于图像变换的双阈值对抗样本检测框架

4.1 图像变换

图像变换是一种通过对图像进行操作和处理的方式, 以改变图像的外观、结构或特征. 在对抗样本检测领域, 图像变换扮演着重要的角色, 许多对抗样本检测方法都是利用图像变换操作来观察对抗样本与良性样本之间的差异, 并进行检测. 本文中使用到的图像变换方法有以下 9 种.

(1) 加噪. 在原始图像 x 中添加噪声, 得到加噪后的图像 x'. 常见的噪声类型有高斯噪声、泊松噪声、盐噪声、胡椒噪声、混合噪声和斑点噪声等. 高斯噪声指为原始图像 x 添加具有高斯分布的噪声. 泊松噪声指为图像添加符合泊松分布的噪声类型. 盐噪声指随机将图像的部分像素值变为最大值 1. 胡椒噪声指随机将图像的部分像素值变为 0. 混合噪声是盐噪声与胡椒噪声的组合, 即随机将图像中的部分像素变为最大值 1 或者最小值 0. 斑点噪声指一种均匀分布的噪声, 添加这种噪声的图像看起来粗糙有噪点. 以高斯噪声为例, 其形式化表达如公式 (6) 所示:

$$x' = x + N(u, \sigma^2) \tag{6}$$

其中, $N(u,\sigma^2)$ 是均值为 u, 方差为 σ^2 的高斯分布.

(2) 平滑滤波. 常用于减少图像 x 中的噪声和细节, 得到的图像 x' 看起来更加平滑. 常见的平滑滤波方法有高斯滤波、最大值滤波、最小值滤波、中值滤波和均值滤波等. 高斯滤波通过使用高斯核对图像进行加权平均, 能有效去除高频噪声, 同时保留图像的整体结构和边缘信息. 最大值滤波将滤波窗口内的像素值替换为窗口内最大值, 能有效去除胡椒噪声. 最小值滤波将滤波窗口内的像素值替换为窗口内的最小值, 能有效去除亮噪声, 但是会导致图像变暗. 中值滤波将滤波窗口内的像素值替换为窗口内的中值, 能有效去除椒盐噪声和斑点噪声. 均值滤波将滤波窗口内的像素值替换为窗口内平均值, 能有效减少噪声, 但可能导致图像模糊. 以高斯滤波为例, 其形式化表达如公式 (7) 所示:

$$x' = G_\delta * x \tag{7}$$

其中, G_δ 是标准差为 δ 的高斯核, $*$ 表示卷积操作.

(3) 位深度减少. 通过降低图像中像素值的表示精度, 来减少图像的颜色深度或灰度级别. 一般彩色图像有 RGB 三通道, 每个通道的位深度为 8 bit. 如果将位深度降低至 1 bit, 图像将只有黑白两种颜色, 色彩会严重缺失. 位深度减少的具体过程可以分为 3 步. 首先确定每个像素可能的最大值 $npp_{\text{int}} = npp - 1$; 将图像 x 乘以 npp_{int}, 然后四舍五入至最相近的整数得到 x_{int}, 再除以 npp_{int} 得到位深度减少后的 x_{float}. x_{float} 表示位深度减少后的图像.

(4) 去高频. 常见的去高频方法有低通滤波器、频域滤波、小波变换等. 本文采用频域滤波, 通过离散余弦变换将图像从空域转换到频域, 并通过设定好的频域系数对高频信息进行过滤, 去除图像中的高频信息. 给定原始图像 x 和过滤比率 $ratio$, 去高频将 x 中每个元素 x_i 过滤高频特征, 形式化表达如公式 (8) 所示:

$$x' = \frac{decode(compress(encode(x_i \times 255), ratio))}{255} \tag{8}$$

其中, $encode$ 是通过 DCT 将图像转换到频域表示; $compress$ 是按照给定的比率 $ratio$ 减少频域表示的信息量, 对图像进行压缩实现高频过滤; $decode$ 将压缩后的数据转换至空间域. 通过频域编码、压缩和解码的过程, 来减少图像中的高频成分, 有助于图像去噪或压缩. 这些操作均在图像的每个颜色通道上独立进行.

(5) 平移. 将图像沿着水平方向和垂直方向移动, 来改变图像中像素的位置. 对于平移后超出原始图像边界的像素进行裁剪, 空出的部位像素进行填 0 操作.

(6) 翻转. 将图像沿着水平或者垂直方向进行镜像反转. 常见的图像翻转有 3 种形式, 水平翻转、垂直翻转和水平和垂直翻转 (旋转 180°). 水平翻转指将图像中每一行的像素都与其对称位置像素进行交换, 垂直翻转则是将图像中每一列的像素都与其对称位置像素进行交换, 水平和垂直翻转则是先进行水平翻转再进行垂直翻转, 或者先进行垂直翻转再进行水平翻转.

(7) 旋转. 将图像围绕中心点按照一定的角度进行旋转, 来改变图像中对象的位置和方向. 图像的旋转角度一般在−360°到 360°之间.

(8) 错切. 通过斜向拉伸或者压缩图像的一部分来改变其形状. 本文使用水平错切方式, 使其在水平方向发生形状扭曲.

(9) 缩放. 按照一定的比例将图像先放大然后缩小, 最终返回与原图像相同大小的图像.

4.2 预测概率分布差异

为了量化经过图像变换前后的预测概率分布的差异, 我们引入了 KL (Kullback-Leibler) 散度[32]作为衡量预测概率分布差异的指标. KL 散度也称相对熵, 主要用于度量两个概率分布之间的差异或者相似性, 其形式化表达如公式 (9) 所示:

$$D_{\mathrm{KL}}(P \parallel Q) = \sum_{i \in A} P(i) \log \frac{P(i)}{Q(i)} \tag{9}$$

其中, P 和 Q 是两个概率分布, i 是集合 A 中的一个事件, $P(i)$ 表示 P 分布在事件 i 上的概率, $Q(i)$ 表示 Q 分布在事件 i 上的概率. 在信息论中, $D_{\mathrm{KL}}(P \parallel Q)$ 常用来表示使用概率分布 Q 拟合真实分布 P 产生的信息损耗. 我们使用 $D_{\mathrm{KL}}(P \parallel Q)$ 来表示经过图像变换后, 分类模型输出的预测概率分布 Q 对原始图像经过分类模型输出的预测概率分布 P 产生的特征损失. KL 散度值越大, 说明变换前后图像的特征分布变化越大; KL 散度值越小, 表示变换前后图像的特征分布变化越小.

4.3 阈值选取

在对抗样本检测中, 阈值的选择一直是一个重要问题, 它直接影响到检测方法的性能和效果. 选取合适的阈值能够有效平衡误报率和漏报率. 我们采用双阈值检测, 比单阈值检测方法多一个下阈值. 无论是单阈值检测还是双阈值检测, 其目标一致: 降低检测器的误报率和漏报率, 保证检测器的性能. 因此, 首先介绍一种单阈值选取方法. 在单阈值的选取中, 为了评价一个阈值的优劣, 这里我们提出了一个检测器性能评价指标 S, 其计算方法如公式 (10) 所示:

$$S = TN \times \frac{TP}{TP + FN} \tag{10}$$

其中, TN 表示真反例 (true negative) 即正确识别良性样本的数量, TP 表示真正例 (true positive) 即正确识别对抗样本的数量, FN 表示假反例 (false negative) 即将对抗样本错误识别为良性样本的数量. 评价指标 S 实质上是真反例与真正例率的乘积. 该指标不仅反映了检测方法识别对抗样本的能力, 同时也考虑到了对良性样本的识别能力, 提供了对检测方法性能更全面的评估. 因此, 当设定阈值的评价指标 S 分数越高, 意味着该阈值能够准确地区分良性样本和对抗样本, 保证检测器的误报率和漏报率维持在较低水平.

单阈值选取流程如下.

1) 选择部分对抗样本与良性样本进行图像变换, 计算其变换前后的概率分布距离, 构成阈值候选列表.

2) 初始化最佳阈值 T_0 以及评价指标 S_0.

3) 遍历阈值候选列表, 计算对应阈值的综合评价指标 S.

4) 判断 S 是否大于 S_0, 若 S 大于 S_0, 更新 S_0 和 T_0, 返回第 3) 步; 否则直接返回第 3) 步. 直至遍历完阈值候选列表.

5) 输出最佳阈值 T_0.

在良性样本与强鲁棒性对抗样本组成的数据集中, 双阈值选取执行单阈值选取流程确定下阈值. 本方案采用

二分类查找, 尝试找到一个合适的阈值, 使得检测器在区分良性样本与鲁棒性较强的对抗样本时, 评价指标 S 取得最大值.

接着我们使用混合对抗样本, 即既包含强鲁棒性也包含弱鲁棒性的对抗样本, 以及良性样本开展实验, 选取大于下阈值的概率分布距离构成上阈值候选列表, 计算每个阈值下的假正例率 (false positive rate, *FPR*) 与真正例率 (true positive rate, *TPR*), 绘制 ROC 曲线, 结合 ROC 曲线找到 *FPR* 尽可能小、*TPR* 尽可能大的阈值设定为上阈值. 最后对选定的阈值区间进行细微调整, 使检测器达到最好的检测效果.

5 实验分析

5.1 实验设置

为了证明本文所提方法的检测性能, 我们选择在 2 个数据集和 3 个分类模型上进行测试. 数据集选择 CIFAR-10 数据集和 ImageNet 数据集. CIFAR-10 是一个小型图像分类数据集, 它包含 10 个不同类别的彩色图像, 其中每个类别有 6000 张图像, 每张图像的分辨率为 32×32 像素. ImageNet 数据集是大规模图像分类数据集, 它包含超过 1000 个类别的图像, 总图像数超过 100 万张, 每张图像分辨率较高通常在几百像素以上. 分类模型则是选择使用训练好的 DenseNet 模型、VGG19 模型以及 ConvNeXt 模型. 将 DenseNet 模型应用于 CIFAR-10 数据集的分类任务上, VGG19 模型和 ConvNeXt 模型应用于 ImageNet 数据集上. DenseNet 模型在 CIFAR-10 数据集上的准确度高达 94.84%, VGG19 和 ConvNeXt 模型在 ImageNet 数据集上准确度高达 71.34% 和 82.30%.

关于对抗样本生成, 我们选择了对模型威胁较大的白盒攻击来生成对抗样本. 采用第 2.1 节里介绍的 5 种常见的白盒攻击方法, 随机选取数据集中部分样本进行攻击, 并筛选出攻击成功的对抗样本进行后续实验. 在 FGSM、BIM、PGD 攻击中, 将攻击参数 ε 控制在 0.1、0.2 和 0.3 之间. 在 DeepFool 攻击中, 将攻击步长设置在 0.02、9 和 16. 在 CW 攻击中, 将置信度参数 k 分别控制在 0、0.5、1.0 和 1.5 之间.

图像变换的参数设置如下所述. 添加噪声的类型为高斯噪声, 平滑滤波使用最大值滤波, 位深度减少设置为减少到 7 bit, 去高频频域系数设定为 0.9, 平移设定为平移 1 个像素, 翻转设定为水平翻转, 旋转角度为−15°, 水平错切系数为 0.3, 缩放设定为放大 1.1 倍.

在本实验的威胁模型中, 我们允许攻击者知道目标模型的信息, 包括模型的结构、参数等. 攻击者可以根据目标模型的信息开展白盒攻击, 但不清楚对抗样本检测器的具体细节.

5.2 评价指标

本文采用 *FPR*、*TPR*、ROC (receiver operating characteristic) 曲线和 AUC (area under the curve)[33]等指标来评估本文检测方法的有效性以及与其他检测方法做对比.

(1) *FPR* 是指在所有实际为负例的样本中被错误识别为正例的比例, 在本文中指所有良性样本中被错误判定为对抗样本的比例, 该值越小表示检测方法正确识别良性样本能力越强. *FPR* 计算公式如下:

$$FPR = \frac{FP}{FP + TN} \tag{11}$$

其中, FP 表示假正例 (false positive) 即将良性样本错误识别为对抗样本的数量.

(2) *TPR* 是指在所有实际为正例的样本中被正确识别为正例的比例, 在文中指所有对抗样本中被正确识别为对抗样本的比例, 该值越大表示检测方法识别对抗样本的能力越强. *TPR* 计算公式如下:

$$TPR = \frac{TP}{TP + FN} \tag{12}$$

(3) ROC 曲线和 AUC. ROC 曲线常用于评估二分类模型的性能. 它以 *FPR* 为横轴 *TPR* 为纵轴, 通过对不同的分类阈值计算对应的 *FPR* 与 *TPR*, 将这些点连成线即为 ROC 曲线. ROC 曲线能够直观地反映模型在不同阈值下的性能表现. AUC 是 ROC 曲线下面积, 常用于衡量模型整体的分类性能. AUC 的通常取值在 0.5–1 之间, 越接近 1 则表示模型性能越好. 当 AUC 值为 0.5 时相当于模型分类能力接近于随机猜测, 当 AUC 值为 1 时表示模型分

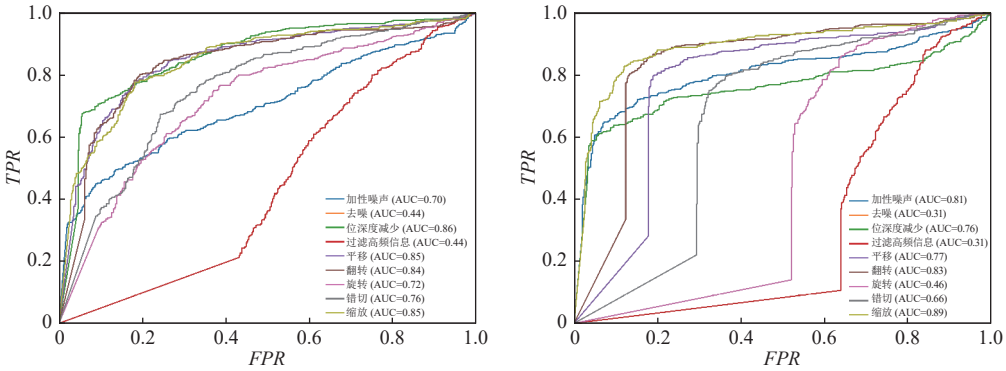
类完全正确.

5.3 实验结果与分析

5.3.1 不同图像变换下检测方法性能

该实验中, 我们采用以上 5 种不同的攻击方法生成对抗样本. 每一类攻击方法在设定的控制参数下生成 50 张对抗样本. 本文的实验设置共 16 中攻击方法和参数的组合, 因此共产生 800 张对抗样本. 对于每一种攻击, 我们随机选择 100 张对抗样本, 及其相同数量的良性样本构建验证集, 以此评估本文所提检测器的性能.

ImageNet 数据集上实验结果如图 4 所示. 从 VGG19 模型的 ROC 曲线可以看到, 本文检测方法对于大部分图像变换均表现出良好的检测效果. 其中, 位深度减少 AUC 值高达 0.86, 说明采用位深度减少能够有效保障 VGG19 抵御对抗样本攻击. 而对于 ConvNeXt 模型, 采用图像缩放能够获得最好的检测效果, AUC 值可达 0.89. 检测性能最差的是平滑滤波和去高频两种变换, 其 AUC 值在 VGG 模型上只有 0.44, 在 ConvNeXt 模型上只有 0.31, 无法识别出对抗样本. 对比两个模型在 ImageNet 数据集上的表现, 我们认为同一个模型对不同图像变换的敏感度不同, 不同的模型对于同一种图像变换的敏感度也不同. 因此会导致同一模型上不同图像变换下检测性能存在一定差异, 同一种图像变换在不同模型上的检测性能也存在差异. 如旋转变换在 VGG19 中 AUC 值为 0.72, 具备较好的对抗样本检测效果. 而其在 ConvNeXt 模型中 AUC 值只有 0.46, 几乎无法准确检测对抗样本.



(a) VGG19 (b) ConvNeXt

图 4 ImageNet 上不同图像变换下检测器的检测性能

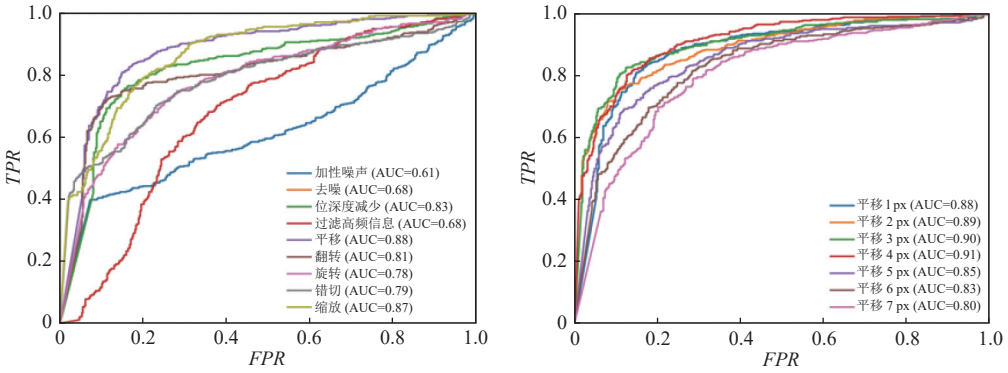
在 CIFAR-10 数据集上实验结果如图 5(a) 所示. 可以看到, 检测性能最好的变换是平移变换, 其 AUC 值最高为 0.88; 检测性能最差的变换是加噪, 其 AUC 值最低为 0.61. 对比两个数据集上实验结果, 我们发现位深度减少、平移和缩放的 AUC 值都较高, 而平滑滤波和去高频的 AUC 值相对较低. 实验结果说明, 对抗样本经过位深度减少、平移和缩放这 3 种变换后, 其弱扰动很轻易被过滤掉, 而强扰动基本无法被去除. 平滑滤波和去高频则无法有效过滤弱扰动, 但可以在一定程度上过滤强扰动这种高频信息. 同时我们观察到, 本方案针对 CIFAR-10 的检测能力要高于对 ImageNet 的检测, 表明本文检测方法对于小尺寸的图像检测效果更优.

5.3.2 不同变换强度下检测方法性能

本文选择 9 种变换中检测效果最好的图像变换, 测试不同变换强度对检测方法性能的影响. 其中, 对于 DenseNet 模型, 我们选择平移变换, 对比其平移 1–7 像素之间的效果, 其 ROC 曲线如图 5(b) 所示; 对于 VGG19 模型, 选择了位深度减少, 对比其位深度减少到 1–7 bit 之间的效果, 其 ROC 曲线如图 6(a) 所示. 对于 ConvNeXt 模型, 选择了缩放, 对比其缩放倍数从 1.1 到 2.3 的效果, 其 ROC 曲线如图 6(b) 所示.

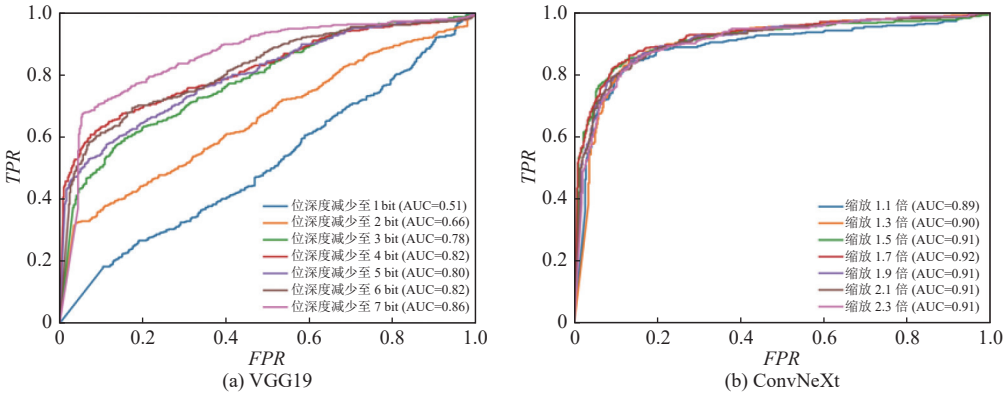
在 ImageNet 数据集上, 本检测器采用位深度减少实现的能够有效检测针对 VGG19 模型的对抗攻击. 当位深度减少至 7 bit 时效果最好, AUC 值高达 0.86; 当位深度减少到 1 bit 时效果最差, AUC 值为 0.51. 我们可以看到, 随着位深度减少的强度增大, 检测器性能也在逐渐下降. 当位深度减少为 1 bit 时, 图像为黑白的二值图像, 无论是良性样本还是对抗样本, 经过变换都产生严重的特征缺失, 导致变换前后模型预测概率分布差异很大, 从而无法有

效进行对抗性检测. 而对于 ConvNeXt 模型, 基于缩放实现的检测器性能最优. 缩放倍数由 1.1 增加到 2.3 时, 检测器的 AUC 值稳定在 0.9 左右. 缩放倍数对检测器在 ConvNeXt 模型的检测性能影响相对较小. 对于 CIFAR-10 数据集, 随着平移像素的增加, 检测器在对 DenseNet 模型的防护中, AUC 值先增加后减少, 在平移 4 像素时达到最大值 0.91, 表现出最强的综合检测能力.



(a) 不同图像变换　　(b) 不同变换强度

图 5　CIFAR-10 上不同变换和不同强度下检测器的检测性能



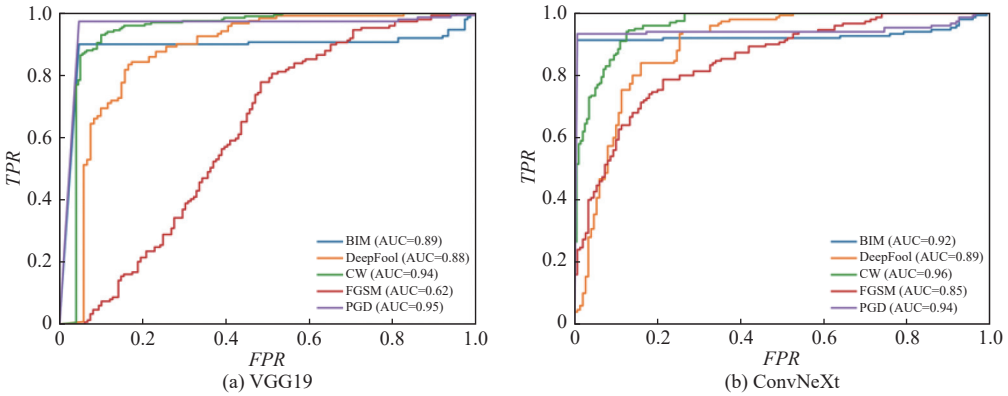
(a) VGG19　　(b) ConvNeXt

图 6　ImageNet 上不同变换强度下检测器的检测性能

5.3.3　不同攻击下检测方法性能

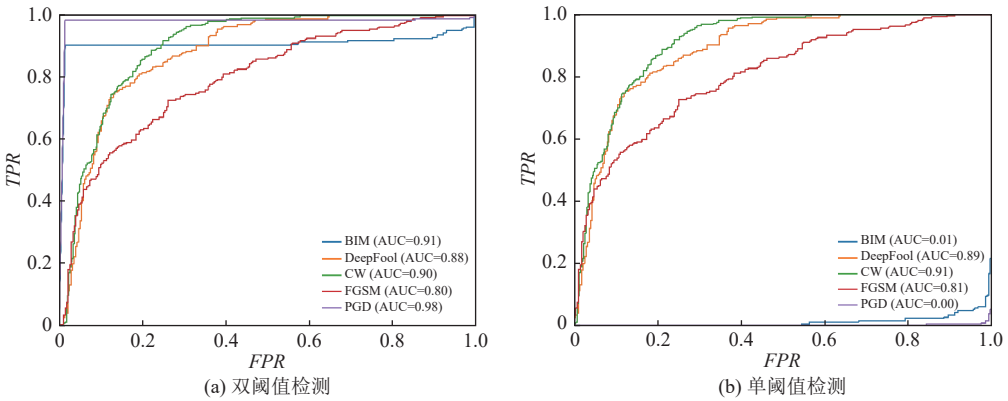
我们选择在最优的条件设置下对不同攻击进行检测. 对于 ImageNet 数据集, 采用位深度减少至 7 来实现本检测器, 保护 VGG19 模型抵御对抗样本攻击; 采用缩放至 1.7 倍来实现本检测器, 保护 ConvNeXt 模型. 对于 CIFAR-10 数据集, 采用位深度减少至 4 位来实现本文所设计的检测器, 来保护 DenseNet 图像分类模型. 检测器在 ImageNet 数据集上的 ROC 曲线如图 7 所示, 在 CIFAR-10 数据集上的 ROC 曲线如图 8(a) 所示.

从图 7 可以看出, 在 ImageNet 数据集上, 本文检测方法对 BIM、DeepFool、CW 和 PGD 攻击表现出良好的检测效果. 其中, 对 PGD 和 CW 攻击, 检测器获得的平均 AUC 值在 0.95 左右. 结果表明, 本文方法可以准确地检测出大部分 PGD 和 CW 对抗样本. 然而, 对于 VGG19 模型中 FGSM 攻击, 检测器获得的 AUC 值只有 0.62, 仍然有部分 FGSM 对抗样本能够规避检测. 在 CIFAR-10 数据集上, 从图 8(a) 可以看到, 面对 PGD 攻击时检测器的 AUC 值达到 0.98, 表明本文方法针对小型图像样本, 能够以非常高效地检测出 PGD 对抗样本. 面对 FGSM 攻击时, 本检测器在 CIFAR-10 数据集上表现最差, 但其 AUC 值仍然高达 0.8. 整体而言, 面对不同类型的对抗攻击, 本文的检测器在 ImageNet 和 CIFAR-10 上均表现出了较强的检测能力.



(a) VGG19 (b) ConvNeXt

图 7 ImageNet 上不同攻击下检测器的检测性能



(a) 双阈值检测 (b) 单阈值检测

图 8 CIFAR-10 上不同攻击下检测器的检测性能

5.3.4 与其他检测方法对比

为了体现本检测方法的优势, 我们将其与先进的检测方法进行对比. 我们选择了 KD+BU[14]、NSS[34]、FS[31]、SFAD[35]和 DNR[36]共 5 种高水平的检测方法, 在 CIFAR-10 数据集上进行比较. 其中, KD+BU 和 NSS 属于有监督检测方法; FS、SFAD 和 DNR 属于无监督检测方法. 这里本文方法的阈值选取在由良性样本与对抗样本组成的数据集中进行, 其中对抗样本由本文所提的 5 类攻击在不同参数下生成, 良性样本与对抗样本数量相同.

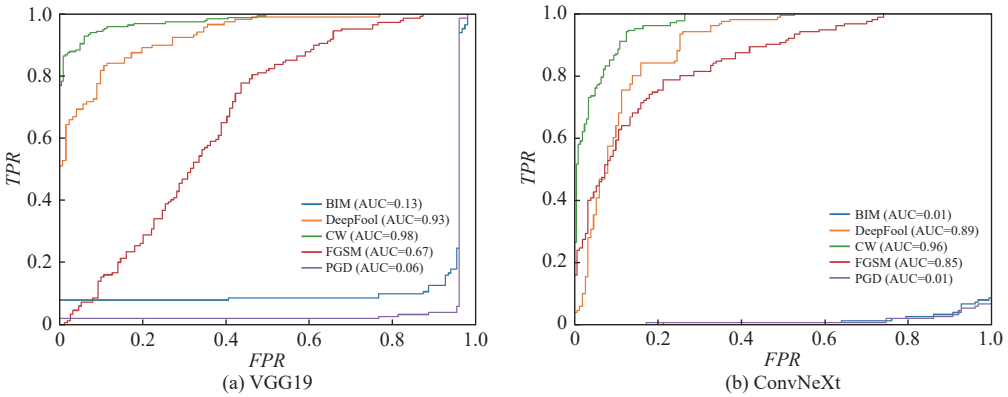
从表 1 中可以看出, 本文方法获得了平均 *TPR* 为 61.99%, 高于其他检测方法. 这说明本文方法在与其他方法相比检测对抗样本的能力最强. 同时, 相比其他方法, 本文方法检测出的对抗样本类型更为全面, 对于 5 种攻击均能有效检测, 针对检测效果最差的 CW 对抗样本, 其检测率也能达到 39.50%. 相比于本文检测方法, 其他方法都存在明显的检测短板. 如 KD+BU 方法对于 CW 和 FGSM 攻击都只能检测 20.00% 到 30.00% 的对抗样本. NSS 方法无法有效检测出 PGD 攻击, 其对 PGD 对抗样本检测率只有 25.61%. FS 方法只能检测出 4.54% 的 BIM 对抗样本和 32.50% 的 FGSM 对抗样本. SFAD 方法也只能检测出 31.74% 的 BIM 对抗样本和 32.81% 的 PGD 对抗样本. DNR 方法只能检测出 10.67% 的 BIM 对抗样本. 虽然我们的方法获得了最高的 *TPR*, 但同时取得了较高的 *FPR*. 这说明本文检测方法存在较高的误报率, 相对于 KD+BU 和 FS 方法更容易产生误报. 但是通过图 8(a) 可以看到, 本文检测方法可以在保证较小的 *FPR* 同时, 对 BIM 及 PGD 两种对抗样本达到良好的检测效果. 而且本文方法不需要更改任何模型或者输入数据的信息, 因此在实际应用中完全可以与其他检测方法结合使用, 以较低的误报率在数据处理阶段筛选出 90.00% 以上的 PGD 与 BIM 对抗样本.

表 1 与其他检测方法在 CIFAR-10 数据集上对比 (%)

检测方法	评价指标	BIM	DeepFool	CW	FGSM	PGD	Average
KD+BU	*FPR*	1.13	1.44	4.94	2.97	4.52	3.00
	TPR	92.01	54.02	29.37	22.78	47.55	49.15
NSS	*FPR*	6.56	6.56	6.56	6.56	6.56	6.56
	TPR	69.95	50.15	44.51	95.73	25.61	57.19
FS	*FPR*	5.07	5.07	5.07	5.07	5.07	5.07
	TPR	4.54	39.18	56.26	32.50	49.68	36.43
SFAD	*FPR*	10.90	10.90	10.90	10.90	10.90	10.90
	TPR	31.74	89.57	59.78	80.14	32.81	58.80
DNR	*FPR*	10.01	10.01	10.01	10.01	10.01	10.01
	TPR	10.67	30.20	35.62	30.23	23.67	26.08
本文	*FPR*	6.67	6.80	6.00	6.71	6.67	6.54
	TPR	88.00	40.82	39.50	40.30	97.33	61.99

5.3.5 消融实验

为了探究下阈值对本文检测方法的影响, 我们使用基于图像变换的单阈值对抗样本检测方法来进行对比实验. 其检测方法的原理是计算输入样本在图像变换前后的预测概率数组之间的距离, 并将该距离与预设的上阈值进行比较. 如果距离大于上阈值, 则被判定为对抗样本; 否则, 判定为良性样本. 该部分实验设置与第 5.3.3 节环境设置相同. 图 8(b) 和图 9 展示了在不同攻击下单阈值检测器的检测性能.



(a) VGG19 (b) ConvNeXt

图 9 ImageNet 上不同攻击下单阈值检测器的检测性能

对比图 7 和图 9 可以发现, 相比双阈值检测方法, 仅设置上阈值的检测方法在 DeepFool、CW 和 FGSM 这 3 种攻击下表现出较强的检测性能. 在 ImageNet 数据集中, 基于双阈值的对抗样本检测方法在保护 VGG19 模型时, 对以上 3 种攻击的 AUC 值分别为 0.88、0.94 和 0.62. 而单阈值检测方法的 AUC 值分别达到 0.93、0.98 和 0.67, 比双阈值方法平均高出约 0.05. 进一步观察发现, 图 9 和图 7 中 VGG19 模型的 ROC 曲线基本保持一致, 最大的区别在于图 7 中, *FPR* 在 0.05 之前对以上 3 种攻击检测率几乎为 0. 这一现象在图 8 中也存在. 这表明仍然存在少部分良性样本在经过图像变换后, 其预测概率分布距离小于下阈值, 被错误判定为对抗样本. 然而, 对于 PGD 和 BIM 这两种对抗攻击, 我们观察到单阈值检测方法在 ImageNet 数据集和 CIFAR-10 数据集上取得的 AUC 值接近 0. 结果表明单阈值检测方法对于这两种攻击完全不具备检测能力. 而双阈值检测方法对于 PGD 和 BIM 攻击表现出优秀的检测性能, 在两个数据集上的 3 个模型中均获得较高的 AUC 值. 这一实验结果说明, 下阈值能有效地检测 PGD、BIM 等生成的强鲁棒性对抗样本. 总而言之, 下阈值的设定虽然会对检测器的误报率有较小的增加, 但能显著降低了检测器的漏报率.

5.3.6 检测效率评估

本文所提的基于图像变换的双阈值对抗样本检测方法提供了一种相对简单的架构, 该架构主要由图像变换组件和特征分布差异计算模块构成. 本检测器对输入样本进行图像变换处理, 通过计算处理前后特征差异来检测对抗样本. 在此过程中, 这种检测框架只要求临时存储变换后的样本结果, 而并不需要为部署检测器保留大量的额外空间.

实验使用 300 张测试样本, 测试了基于图像变换的双阈值检测方法在不同数据集上的推理时间开销, 结果如表 2 所示. 在 ImageNet 数据集上, 图像变换耗时仅为 0.4 s 左右. KL 距离的计算需要执行两次目标模型的推理, 在 VGG19 模型上耗时 26.15 s, 在 ConvNeXt 模型上耗时 17.5 s. 在 CIFAR-10 数据集上, 图像变换部分仅需 0.012 s, KL 距离的计算耗时 4.34 s. 由此观察出, 在该检测方法运行时间中, 主要的时间开销发生在 KL 距离计算上. 这是因为在执行距离计算之前, 需要利用目标图像分类模型对变换前后的图像进行推理. 我们还注意到, CIFAR-10 数据集推理效率高于 ImageNet 数据集. 这主要得益于 CIFAR-10 图像尺寸较小, 模型执行推理速度更快, 使得检测器的运行更为高效. 从整体的检测效率来看, 检测器在 ImageNet 上对每张样本的平均检测时间低于 0.09 s, 在 CIFAR-10 上仅为 0.015 s, 结果表明该检测器具有较高的效率. 总体而言, 基于图像变换的双阈值检测方法拥有计算效率高和资源占用少的优势, 适用于对延迟敏感和资源紧张的应用场景.

表 2 检测架构推理效率

数据集	分类模型	样本数量	图像变换时间 (s)	KL计算时间 (s)	整体推理时间 (s)	平均推理时间 (s)
ImageNet	VGG19	300	0.43	26.15	26.58	0.089
	ConvNeXt	300	0.38	17.50	17.93	0.06
CIFAR-10	DenseNet	300	0.012	4.34	4.36	0.015

6 总 结

针对现有对抗样本检测方法在输入数据变换处理的特征差异性辨识上存在的局限性, 本文提出了一种创新的基于图像变换的双阈值对抗样本检测方法. 该方法首先执行图像变换处理, 继而量化变换前后输入样本的预测概率分布距离. 通过设定两个阈值定义的区间, 此方法可判别上述距离是否落在该区间内, 以此来检测对抗样本. 在 ImageNet 和 CIFAR-10 数据集上, 我们对不同的模型采取不同的图像变换及强度对 5 种对抗样本进行了测试. 实验结果表明, 相比于当前的单阈值检测方案, 本文所提的检测方法基于不同图像变换的检测效果均得到了明显提升. 具体而言, 在 ImageNet 数据集上使用缩放变换, 对于不同模型均表现出良好的检测能力, 而针对 CIFAR-10 这种小尺寸图像的数据集, 使用平移变换的检测效果最优. 在针对 PGD 产生的对抗样本检测上, 我们能够在保持低误报率的同时, 实现 90% 以上的高检测率. 与其他检测方案相比较, 本文方法保持了较高的检测精度, 同时具备较强的泛化能力. 下阈值的设定能够显著提升检测器的泛化能力, 提升了对高鲁棒性对抗样本的检测性能. 然而, 下阈值潜在地限制了良性样本的识别边界, 导致少量对图像变换稳定的良性样本被错误地判定为对抗样本, 造成一定的误报率. 在未来的研究中, 我们将充分发挥双阈值检测兼容性强的优势, 考虑同时使用多种变换方法或者与其他检测方法联合, 降低对抗样本检测的误报率同时提升其准确度.

References:

[1] Papa L, Russo P, Amerini I, Zhou LP. A survey on efficient vision transformers: Algorithms, techniques, and performance benchmarking. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2024, 46(12): 7682–7700. [doi: 10.1109/TPAMI.2024.3392941]

[2] Cheng XZ, Jin T, Li LJ, Lin W, Duan XY, Zhao Z. OpenSR: Open-modality speech recognition via maintaining multi-modality alignment. In: Proc. of the 61st Annual Meeting of the Association for Computational Linguistics (Vol. 1: Long Papers). Toronto: Association for Computational Linguistics, 2023. 6592–6607. [doi: 10.18653/v1/2023.acl-long.363]

[3] Min B, Ross H, Sulem E, Pouran Ben Veysel A, Nguyen TH, Sainz O, Agirre E, Heintz I, Roth D. Recent advances in natural language processing via large pre-trained language models: A survey. ACM Computing Surveys, 2023, 56(2): 30. [doi: 10.1145/3605943]

[4] Yu JL, Yin HZ, Xia X, Chen T, Li JD, Huang Z. Self-supervised learning for recommender systems: A survey. IEEE Trans. on

Knowledge and Data Engineering, 2024, 36(1): 335–355. [doi: 10.1109/TKDE.2023.3282907]

[5] Szegedy C, Zaremba W, Sutskever I, Bruna J, Erhan D, Goodfellow I, Fergus R. Intriguing properties of neural networks. arXiv: 1312.6199, 2014.

[6] Mao JG, Shi SS, Wang XG, Li HS. 3D object detection for autonomous driving: A comprehensive survey. Int'l Journal of Computer Vision, 2023, 131(8): 1909–1963. [doi: 10.1007/s11263-023-01790-1]

[7] Nguyen K, Proença H, Alonso-Fernandez F. Deep learning for iris recognition: A survey. ACM Computing Surveys, 2024, 56(9): 223. [doi: 10.1145/3651306]

[8] Sun SH, Goldgof GM, Butte A, Alaa AM. Aligning synthetic medical images with clinical knowledge using human feedback. In: Proc. of the 37th Int'l Conf. on Neural Information Processing Systems. New Orleans: Curran Associates Inc., 2023. 13408–13428.

[9] Gong ZT, Wang WL. Adversarial and clean data are not twins. In: Proc. of the 6th Int'l Workshop on Exploiting Artificial Intelligence Techniques for Data Management. Seattle: ACM, 2023. 6. [doi: 10.1145/3593078.3593935]

[10] Papernot N, McDaniel P, Wu X, Jha S, Swami A. Distillation as a defense to adversarial perturbations against deep neural networks. In: Proc. of the 2016 IEEE Symp. on Security and Privacy. San Jose: IEEE, 2016. 582–597. [doi: 10.1109/SP.2016.41]

[11] Liang B, Li HC, Su MQ, Li XR, Shi WC, Wang XF. Detecting adversarial image examples in deep neural networks with adaptive noise reduction. IEEE Trans. on Dependable and Secure Computing, 2021, 18(1): 72–85. [doi: 10.1109/TDSC.2018.2874243]

[12] Zhou T, Gan R, Xu DW, Wang JY, Xuan Q. Survey on adversarial example detection of images. Ruan Jian Xue Bao/Journal of Software, 2024, 35(1): 185–219 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/6834.htm [doi: 10.13328/j.cnki.jos.006834]

[13] Hendrycks D, Gimpel K. Early methods for detecting adversarial images. arXiv:1608.00530, 2017.

[14] Feinman R, Curtin RR, Shintre S, Gardner AB. Detecting adversarial samples from artifacts. arXiv:1703.00410, 2017.

[15] Lust J, Condurache AP. GraN: An efficient gradient-norm based detector for adversarial and misclassified examples. arXiv:2004.09179, 2020.

[16] Liu H, Zhao B, Guo JB, Zhang KH, Liu P. A lightweight unsupervised adversarial detector based on autoencoder and isolation forest. Pattern Recognition, 2024, 147: 110127. [doi: 10.1016/j.patcog.2023.110127]

[17] Wang YC, Li XG, Yang L, Ma JF, Li H. ADDITION: Detecting adversarial examples with image-dependent noise reduction. IEEE Trans. on Dependable and Secure Computing, 2024, 21(3): 1139–1154. [doi: 10.1109/TDSC.2023.3269012]

[18] Zheng ZH, Hong PY. Robust detection of adversarial attacks by modeling the intrinsic properties of deep neural networks. In: Proc. of the 32nd Int'l Conf. on Neural Information Processing Systems. Montréal: Curran Associates Inc., 2018. 7924–7933.

[19] Eniser HF, Christakis M, Wüstholz V. RAID: Randomized adversarial-input detection for neural networks. arXiv:2002.02776, 2020.

[20] Ma SQ, Liu YQ, Tao GH, Lee WC, Zhang XY. NIC: Detecting adversarial samples with neural network invariant checking. In: Proc. of the 2019 Network and Distributed System Security Symp. San Diego, 2019. 1–15. [doi: 10.14722/ndss.2019.23415]

[21] Tian SX, Yang GL, Cai Y. Detecting adversarial examples through image transformation. In: Proc. of the 32nd AAAI Conf. on Artificial Intelligence. New Orleans: AAAI Press, 2018. 4139–4146. [doi: 10.1609/aaai.v32i1.11828]

[22] Ryu G, Choi D. Detection of adversarial attacks based on differences in image entropy. Int'l Journal of Information Security, 2024, 23(1): 299–314. [doi: 10.1007/s10207-023-00735-6]

[23] Goodfellow IJ, Shlens J, Szegedy C. Explaining and harnessing adversarial examples. arXiv:1412.6572, 2015.

[24] Kurakin A, Goodfellow I, Bengio S. Adversarial examples in the physical world. arXiv:1607.02533, 2017.

[25] Madry A, Makelov A, Schmidt L, Tsipras D, Vladu A. Towards deep learning models resistant to adversarial attacks. arXiv:1706.06083, 2019.

[26] Moosavi-Dezfooli SM, Fawzi A, Frossard P. DeepFool: A simple and accurate method to fool deep neural networks. In: Proc. of the 2016 IEEE Conf. on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 2574–2582. [doi: 10.1109/CVPR.2016.282]

[27] Carlini N, Wagner D. Adversarial examples are not easily detected: Bypassing ten detection methods. In: Proc. of the 10th ACM Workshop on Artificial Intelligence and Security. Dallas: ACM, 2017. 3–14. [doi: 10.1145/3128572.3140444]

[28] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556, 2015.

[29] Huang G, Liu Z, Van der Maaten L, Weinberger KQ. Densely connected convolutional networks. In: Proc. of the 2017 IEEE Conf. on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017. 2261–2269. [doi: 10.1109/CVPR.2017.243]

[30] Liu Z, Mao HZ, Wu CY, Feichtenhofer C, Darrell T, Xie SN. A ConvNet for the 2020s. In: Proc. of the 2022 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022. 11966–11976. [doi: 10.1109/CVPR52688.2022.01167]

[31] Xu WL, Evans D, Qi YJ. Feature squeezing: Detecting adversarial examples in deep neural networks. In: Proc. of the 2018 Network and Distributed System Security Symp. San Diego, 2018. 1–15. [doi: 10.14722/ndss.2018.23198]

[32] Cui JQ, Tian ZT, Zhong ZS, Qi XJ, Yu B, Zhang HW. Decoupled Kullback-Leibler divergence loss. arXiv:2305.13948, 2024.

[33] Zhou ZY, Dou WS, Li S, Kang LY, Wang S, Liu J, Ye D. DiTing: Semi-supervised adversarial training approach for robust out-of-distribution detection. Ruan Jian Xue Bao/Journal of Software, 2024, 35(6): 2936–2950 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/6928.htm [doi: 10.13328/j.cnki.jos.006928]

[34] Kherchouche A, Fezza SA, Hamidouche W, Déforges O. Detection of adversarial examples in deep neural networks with natural scene statistics. In: Proc. of the 2020 Int'l Joint Conf. on Neural Networks. Glasgow: IEEE, 2020. 1–7. [doi: 10.1109/IJCNN48605.2020.9206959]

[35] Aldahdooh A, Hamidouche W, Déforges O. Revisiting model's uncertainty and confidences for adversarial example detection. Applied Intelligence, 2023, 53(1): 509–531. [doi: 10.1007/s10489-022-03373-y]

[36] Sotgiu A, Demontis A, Melis M, Biggio B, Fumera G, Feng XY, Roli F. Deep neural rejection against adversarial examples. EURASIP Journal on Information Security, 2020, 2020: 5. [doi: 10.1186/s13635-020-00105-y]

附中文参考文献:

[12] 周涛, 甘燃, 徐东伟, 王竟亦, 宣琦. 图像对抗样本检测综述. 软件学报, 2024, 35(1): 185–219. http://www.jos.org.cn/1000-9825/6834.htm [doi: 10.13328/j.cnki.jos.006834]

[33] 周志阳, 窦文生, 李硕, 亢良伊, 王帅, 刘杰, 叶丹. 谛听: 面向鲁棒分布外样本检测的半监督对抗训练方法. 软件学报, 2024, 35(6): 2936–2950. http://www.jos.org.cn/1000-9825/6928.htm [doi: 10.13328/j.cnki.jos.006928]

刘会(1992－), 男, 博士, CCF 专业会员, 主要研究领域为人工智能安全, 可信计算, 隐私保护, 大语言模型.

王敬华(1965－), 男, 副教授, CCF 专业会员, 主要研究领域为人工智能安全, 数据库, 数据挖掘.

文福举(2000－), 男, 硕士生, 主要研究领域为人工智能安全, 隐私保护, 大语言模型.

赵波(1972－), 男, 博士, 教授, 博士生导师, CCF 高级会员, 主要研究领域为可信计算, 人工智能.

杜红琴(2001－), 女, 硕士生, 主要研究领域为人工智能安全, 隐私保护, 大语言模型.