

车牌识别系统中基于特征不变量的对抗检测

朱孝羽, 唐鹏, 张浩臣, 邱卫东, 黄征

(上海交通大学网络空间安全学院, 上海 200240)

摘要: 深度神经网络已成为人们日常生活的重要部分。但研究人员发现, 深度神经网络容易受到对抗样本的威胁, 使网络模型发生错误分类等异常行为。对抗样本的存在极大地威胁了深度学习的应用, 特别是在对于安全性需求非常敏感的场景, 如车牌识别系统。目前, 大部分现有的对抗样本防御和检测技术对某些特定类型的对抗攻击可以呈现出很好的效果, 但通常情况下, 它们并不能对所有类型的对抗攻击都具有通用性。针对真实场景下车牌识别系统的对抗样本攻击, 通过分析神经网络在干净样本上训练的内在变化和干净样本之间的维度复杂度, 提出了基于神经网络不变量和局部固有维数不变量的特征不变量无监督对抗样本检测系统, 称为特征不变量对抗检测 (feature invariant adversarial detection, FIAD)。该研究将本检测系统部署到广泛使用的开源车牌识别系统HyperLPR和EasyPR中, 并以真实的中国车牌数据集 (Chinese city parking dataset, CCPD) 对该检测系统进行了广泛的实验。对11种不同类型的攻击进行实验, 结果表明, 相比其他4种先进的检测方法, FIAD可以在更低的假阳率下有效地检测所有攻击, 且平均检测准确率始终高达99%。因此, FIAD对各种类型的对抗攻击具有良好的通用性。

关键词: 深度神经网络; 对抗样本检测; 车牌识别; 特征不变量

中图分类号: TP393

文献标志码: A

DOI: 10.11959/j.issn.2096-109x.2024080

Adversarial detection based on feature invariant in license plate recognition systems

ZHU Xiaoyu, TANG Peng, ZHANG Haochen, QIU Weidong, HUANG Zheng

School of Cyber Science and Engineering, Shanghai Jiao Tong University, Shanghai 200240, China

Abstract: Deep neural networks have become an integral part of people's daily lives. However, researchers observed that these networks were susceptible to threats from adversarial samples, leading to abnormal behaviors such as misclassification by the network model. The presence of adversarial samples poses a significant threat to the application of deep neural networks, especially in security-sensitive scenarios like license plate recognition systems. Presently, most existing defense and detection technologies against adversarial samples show promising results for

收稿日期: 2024-01-20; 修回日期: 2024-03-19

通信作者: 朱孝羽, zhuxiaoyu@sjtu.edu.cn

基金项目: 国家重点研发计划资助 (2023YFB3106501)

Foundation Item: The National Key R&D Program of China (2023YFB3106501)

引用格式: 朱孝羽, 唐鹏, 张浩臣, 等. 车牌识别系统中基于特征不变量的对抗检测[J]. 网络与信息安全学报, 2024, 10(6): 59-70.

Citation Format: ZHU X Y, TANG P, ZHANG H C, et al. Adversarial detection based on feature invariant in license plate recognition systems[J]. Chinese Journal of Network and Information Security, 2024, 10(6): 59-70.

specific types of adversarial attacks. However, they often lack generality in addressing all types of adversarial attacks. In response to adversarial sample attacks on real-world license plate recognition systems, an unsupervised adversarial sample detection system named FIAD was proposed, which was based on analyzing the inherent variations in neural networks trained on clean samples and the dimensional complexity between clean samples. FIAD utilized neural network invariants and local intrinsic dimensionality invariants for effective sample detection. The detection system was deployed into widely used open-source license plate recognition systems, HyperLPR and EasyPR, and extensive experiments were conducted using the real dataset CCPD. The results from experiments involving 11 different types of attacks indicate that, compared to 4 other advanced detection methods, FIAD can effectively detect all these attacks at a lower false positive rate, with an accuracy consistently reaching 99%. Therefore, FIAD exhibits good generality against various types of adversarial attacks.

Keywords: deep neural network, adversarial sample detection, license plate recognition, feature invariants

0 引言

近年来,深度学习在众多领域取得了一定的成果,如图像分割^[1]、语音识别^[2]、自动驾驶和物体追踪^[3]等。然而,研究人员发现深度神经网络(deep neural network, DNN)易受对抗样本的攻击^[4]。这些对抗样本在感官上看起来与原始干净图像一致,但却会使DNN模型输出错误的分类结果。

目前,许多安全关键型的应用程序正在将DNN集成为其系统的一部分。例如,许多停车场使用基于DNN的车牌识别系统(如HyperLPR^[5]、EasyPR^[6])取代人工登记。然而,对抗样本的存在对这些应用构成了重大威胁,被攻击的深度神经网络会导致严重的后果。在车牌识别系统中,攻击者同样可以对车牌添加干扰,使模型输出错误的车牌号,从而逃脱交通排查和躲避监控^[7]。因此,对抗样本的检测和防御对于深度学习模型在实际中的应用安全具有重大意义。

现有对抗样本的工作包括对抗样本的防御和检测技术^[8]。防御技术如对抗性训练^[9]、特征去噪^[10-11]、预处理和梯度掩蔽^[12]等,试图使用各种方法来强化DNN,从而使DNN模型对抗攻击更为坚固,以正确地分类对抗样本。但是,攻击者针对这种防御技术开发了更为有效的攻击技术,如基于Jacobian的显著性图攻击(jacobian-based saliency map attack, JSMA)^[13]。与防御技术不同,对抗样本的检测技术,如基于统计的局部固有维数^[14]、基于一致性的特征压缩^[15]和网络不变量^[16]等是在将样本传递给DNN之前对其进行检测。在车牌识别等安全性要求较高、环境因素复杂多变的场景下,

相比强化DNN从而获得正确分类的防御技术,及时检测对抗样本并发出警告显得更为重要。

本文针对真实场景下车牌识别系统的对抗样本攻击,通过分析神经网络在干净样本中训练的内在变化和干净样本之间的维度复杂度,提出了基于神经网络不变量和局部固有维数(local intrinsic dimensionality, LID)不变量的特征不变量无监督对抗样本检测系统,称为特征不变量对抗检测(feature invariant adversarial detection, FIAD)。具体来说,针对每类干净样本,FIAD分别构建神经网络每层激活值不变量模型和每层LID不变量模型,以标记干净样本的正确分布特点。在实际运行中,对于给定的样本,若其偏离干净样本不变量的分布状态,则视为对抗样本。因此,FIAD属于只需干净样本即可构建检测系统的无监督检测方法。在车牌识别场景下,FIAD的无监督性也使本文无须考虑复杂的自然环境对检测系统带来的各种挑战。在真实的中国车牌数据集(Chinese city parking dataset, CCPD)^[17]和车牌识别系统上的广泛实验表明,相比其他4种先进的检测方法,FIAD能有效地检测CCPD中的对抗攻击,实现更高的检测准确率和更低的假阳率,准确率始终高达99%。综上所述,本文的主要创新点如下。

1) 本文提出了针对车牌识别系统的基于特征不变量的对抗样本检测方法,包括提取干净样本的神经网络每层激活值不变量和干净样本之间的每层局部固有维数不变量。

2) FIAD是对每类标签单独构建对抗检测模型的无监督方法。相比现有所有样本统一检测的思路,FIAD的训练时间更短,检测效率更高。

3) 广泛的实验结果表明, FIAD 能检测出主流的 11 种对抗攻击, 且检测准确率始终高达 99%。与其他 4 种先进的检测方法相比, FIAD 有更好的检测效果。

4) 本文通过设计白盒自适应攻击来评估 FIAD 的鲁棒性。实验结果表明, FIAD 显著提高了攻击者构建对抗样本的难度, 对抗样本在 CCPD 上的平均 L_2 范数增大了 1.85 倍。

1 背景和相关工作

1.1 对抗样本攻击方法

深度神经网络易受对抗样本的攻击^[4]。对抗样本与能够被正确分类的样本(即干净样本)在直观上非常相似, 但 DNN 会对这两个输入产生不同的预测输出^[8], 如图 1 所示。现有的对抗攻击一般以两种不同的方式来生成对抗样本, 即基于梯度和基于内容的攻击方式^[13]。

1) 基于梯度的攻击。攻击者将生成对抗样本视为优化问题。假设 x 和 x' 分别为干净样本和对抗样本, 且其在模型 M 中的预测输出分别为 $C(x)$ 和 $C(x')$, 则生成对抗样本可视为以下优化问题:

$$\min \Delta(x, x') \text{ s.t. } C(x') \neq C(x) \wedge x' \in X \quad (1)$$

其中, $\Delta(\cdot)$ 为距离函数, X 为输入空间。在求解该问题时, 基于梯度的攻击将采用梯度引导的最优搜索法, 并通常使用 L_p 范数作为 $\Delta(\cdot)$, 如 L_0 、 L_2 和 L_∞ 范数^[18]。

目前这类攻击的主流方法包括快速梯度符号法

(fast gradient sign method, FGSM)^[7]、基本迭代法(basic iterative method, BIM)^[19]、CW 攻击(carlini and wagner attack)^[18]和 DeepFool^[20]等。CW 攻击根据使用 L_p 范数的不同, 分为 3 种不同的攻击方式, 即使使用 L_0 的 CW_0 、使用 L_2 范数的 CW_2 和使用 L_∞ 的 CW_∞ 。

2) 基于内容的攻击。攻击者会制作与真实干净样本在语义内容上一致的输入补丁, 以模拟现实中的场景, 如图像水印和相机镜头上的灰尘引起的黑点^[21]。因此, 攻击者有意地在图像中添加补丁对此进行模拟, 从而干扰输入以实现欺骗 DNN 的效果。但是, 这种对抗样本往往是人类可识别的, 因为其具有较大的 $\Delta(\cdot)$ 距离^[22]。

基于内容的攻击包括 DeepXplore^[21]和基于 NSGA-II 的黑盒物理攻击^[23]等。DeepXplore 利用语义信息来执行扰动, 并提出了 3 种攻击方式: 光照攻击(light)、矩形修补(rectangle patch, RP)和污垢攻击(dirt)。此外, NSGA-II 是带精英策略的非支配遗传算法。攻击者将添加的扰动设为黑块, 模拟车牌上的淤泥或镜头上的污渍, 并将其视为遗传算法中的染色体基因, 从而进行车牌的种群进化, 以生成对抗样本。这些攻击在 CCPD 上的对抗样本如图 1 所示。

1.2 对抗样本防御方法

防御方法的目标是强化 DNN 模型^[8]。对抗性训练方法^[9]在 DNN 的训练中包含了对抗样本, 如加入 PGD 攻击生成的对抗样本^[9]。特征去噪^[10-11]是指在模型推理时, 对输入样本的特征进行去

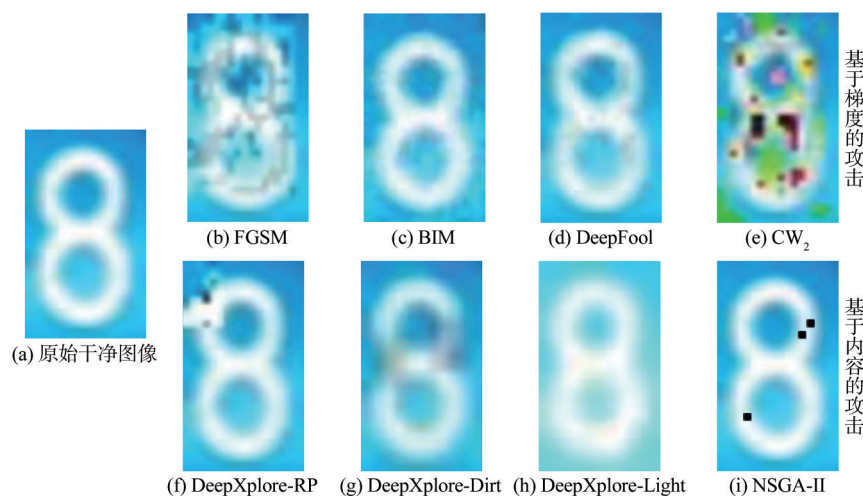


图1 不同攻击在 CCPD 上的对抗样本
Figure 1 Adversarial samples of different attacks on the CCPD

噪, 如 HGD (high-level guided denoiser)^[11]。预处理^[24]是对输入的样本去除扰动噪声后, 再将去噪后的输入传递给 DNN, 如像素偏移算法和超分辨率技术^[24]。梯度掩蔽^[12]试图训练 DNN 使其梯度接近于 0, 从而使 DNN 模型对输入样本中的小扰动不太敏感, 如防御性知识蒸馏策略。

1.3 对抗样本检测方法

检测方法的目标是确定输入是否为对抗样本。主流的检测方法可分为 4 类: 辅助模型、统计测量、特征压缩和网络不变性^[8]。

1) 辅助模型。通过监测干净样本和对抗样本的行为以提取特定的特征从而构建分类器。文献[25]基于自然场景统计 (natural scene statistics, NSS) 构建了以对比度特征为划分标准的二元分类器。Erase&Restore^[26]以干净样本和 L_2 范数对抗样本为数据集训练得到分类器模型。此外, 还有基于阈值和选择分类器的联合检测方法 SFAD (selective and feature based adversarial detection)^[27]。

2) 统计测量。该方法计算干净样本和对抗样本的不同统计特性, 然后利用此特性来构建检测器, 这些特性大多与训练数据的分布相关。LID (local intrinsic dimensionality)^[14]常被用来计算输入样本与其邻居的距离分布, 以评估围绕该样本的区域空间维度。此类方法还包括 PCA (principal component analysis)^[28]、 k -NN (k -nearest neighbors)^[29]等。

3) 特征压缩。该方法旨在压缩输入样本中不必要的特征, 以限制攻击者可用的特征空间。常见的特征压缩 (feature squeezing, FS) 包括位深度^[15]和自适应噪声降低^[30]。若原始输入与压缩后输入的模型预测之间超过一定的识别阈值, 则认为该输入为对抗样本。

4) 网络不变性。近年来, 研究人员发现干净样本和对抗样本会产生不同的特征图和网络层的不同激活值。因此, 分析网络不变性是许多检测方法的核心。RAID (randomized adversarial-input detection)^[31]以干净样本和对抗样本之间神经元激活值的差异为输入, 构建了二元分类器。NIC (neural-network invariant checking)^[16]为 DNN 的每层都构建一组检测模型来描述干净样本的激活值来源和分布, 并以此为标准使用一类分类器检测对抗样本。目前, 基于网络不变性思

想的检测方法较少, 因此这在未来将是很有发展潜力的研究方向。

此外, 还有一些基于去噪思想的检测方法 (如 MagNet^[32]) 和基于对象的检测方法 (如 UnMask^[33])。

1.4 威胁模型

假设攻击者已知有关原始识别模型的所有信息, 包括模型的结构和权重, 以便构造出针对原系统的对抗样本。但是, 检测方法不知道用于生成对抗样本的具体攻击方式。根据攻击者对检测方法的了解程度, 可分为以下两种情况。

1) 攻击者对检测方法一无所知, 即攻击者仅基于原始识别模型来生成对抗样本。

2) 攻击者对检测方法有全面的了解。由于检测方法本身也属于机器学习, 也会受到对抗样本的攻击^[34]。同样, 现有的检测技术也会遇到相同的对抗样本的攻击问题^[16]。值得注意的是, 神经网络模型都属于近似模型, 因此无法完全避免对抗样本的攻击, 但本文的目标是大幅度提高攻击者的攻击难度。

本文对两种威胁模型分别展开了实验评估。结果表明, 本文的 FIAD 可以成功检测实验中的所有攻击, 同时大幅度提高自适应白盒攻击的攻击难度。

2 基于特征不变量的对抗样本检测

2.1 神经网络激活值分布不变量

通过观察各种攻击下 DNN 的内部结构, 研究人员发现攻击者主要扰乱了 DNN 中的激活值分布不变量^[16]。具体来说, 若神经元的激活函数 (如 ReLU) 返回非零值, 则认为该神经元被激活。DNN 的每层可视为一种函数计算, 该函数获取上层激活神经元的激活值, 并利用本层的权重执行乘法运算, 然后应用本层的激活函数来确定本层激活的神经元。因此, 前层的激活神经元决定了后续层中神经元的激活状态。目前很多攻击是基于改变激活值分布关系来生成对抗样本。

例如, 针对一个三分类的全连接网络, 图 2 展示了输入不同样本时该网络的激活值分布状态。彩色为被激活的神经元, 白色为未被激活的神经元, 颜色越深代表激活值越大。对于干净的 A 类样本, Lay_1 的神经元 1、2、3 和 4 导致了

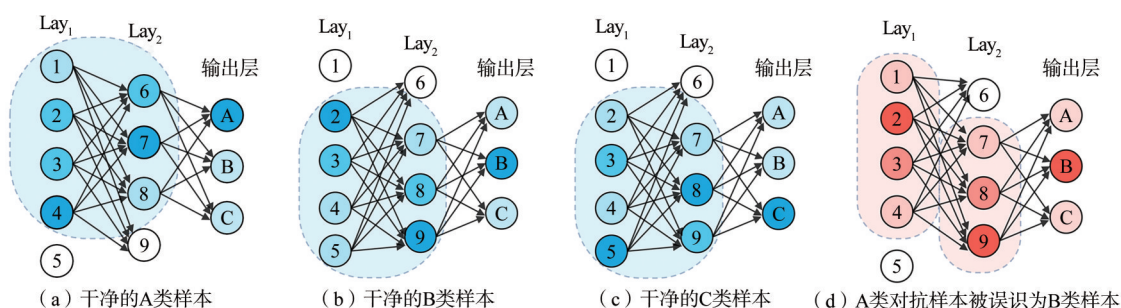


图2 不同样本的激活值分布
Figure 2 Activation value distribution of different samples

Lay₂的6、7和8。对于干净的B类样本，Lay₁的2、3、4和5导致了Lay₂的7、8和9。因此，不同标签的干净样本的每层激活值来源分布存在明显不同。干净的C类样本的激活值来源分布与干净的B类样本相同，但其每层的激活值大小分布与B类样本不同，如C类样本的神经元5和8的激活值明显大于B类样本，从而导致了C类样本的输出概率大于B类样本，使模型的预测结果为C类样本。因此，不同标签的干净样本的每层激活值大小也存在明显不同。

如图2(d)所示，对于被误识为B类的A类对抗样本，Lay₁中的1、2、3和4导致Lay₂的7、8和9，最终被误识为B类样本。显然，此样本与干净B类样本的神经元激活值分布不同，这通常是神经网络内部的不稳定性所致，即前层较小的变化会导致网络后层产生一组完全不同的激活值[16]。在图2(d)中，模型误识的根本原因在于该模型对神经元4的变化非常敏感。

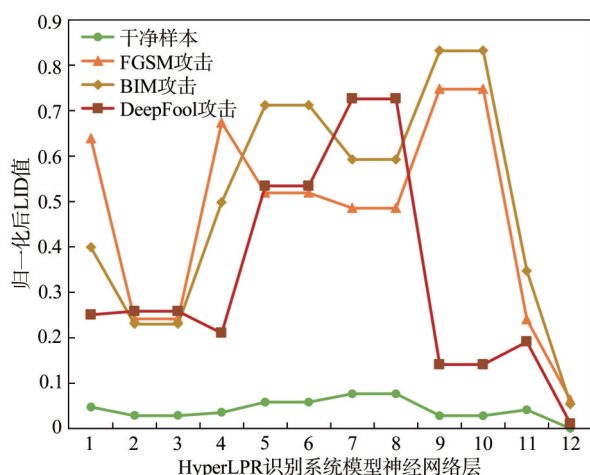


Figure 3 Comparison of LID values between a clean sample and its adversarial sample in CCPD

综上所述，给定与A类样本 x 相似的对抗样本 x' ，若模型将 x' 误识为B类样本，则 x' 通常与干净的A类或B类样本有不同的激活值分布状态。

2.2 局部固有维数不变量

局部固有维数是衡量数据复杂性的指标，用于表征数据的分布空间复杂度。DNN模型中的LID常用来量化模型在输入空间中某点附近的敏感度[14]。

如图3所示，对抗样本的LID明显高于干净样本，且干净样本在DNN每层的LID相对稳定。因此，LID能有效地捕获样本空间中的内在维数特性。本文从体积表达式出发得到LID近似估计表达式。

三维空间体积表达式有 $V \propto r^3$ ，其中， r 为空间中心点到外围的某种距离，可视为半径。对于一般性的欧几里得空间则存在 $V \propto r^m$ ，其中， m 代表数据所在空间的维数，即LID要表征的数据特征。因此，可得 m 关于体积 V 和样本半径 r 的表达式：

$$\frac{V_2}{V_1} = \left(\frac{r_2}{r_1}\right)^m \Rightarrow m = \frac{\ln(V_2/V_1)}{\ln(r_2/r_1)} \quad (2)$$

设 $F(r)$ 为样本空间中以某样本 x 为中心， r 为半径的空间体积 $V = F(r)$ 。为求半径 r 的维数 m 即LID，有极限形式的半径 $r_2 = (1 + \varepsilon)r_1$ ， $r_1 = r$ ， $\varepsilon \rightarrow 0$ ，进而得LID的定义为：

$$\text{LID}_F(r) \triangleq \lim_{\varepsilon \rightarrow 0} \frac{\ln(F((1 + \varepsilon) \cdot r)/F(r))}{\ln(1 + \varepsilon)} = \frac{r \cdot F'(r)}{F(r)} \quad (3)$$

取 r 的极限，即为样本 x 处的LID：

$$\text{LID}_F = \lim_{r \rightarrow 0} \text{LID}_F(r) \quad (4)$$

在实际计算中，必须从本地数据样本中估算LID，其中最大似然估计[35]在统计效率和复杂性

之间表现出较好的效果。样本 x 的LID最大似然估计可表示为:

$$\text{LID}(x) \approx -\left(\frac{1}{k} \sum_{i=1}^k \log \frac{r_i(x)}{r_{\max}(x)}\right)^{-1} \quad (5)$$

其中, k 是邻居的数量, $r_i(x)$ 和 $r_{\max}(x)$ 分别为 x 到第 i 个邻居的距离和最远邻居的距离。

2.3 对抗检测方案的设计

2.3.1 方案概述

图4展示了车牌识别场景下基于特征不变量的对抗样本检测FIAD系统框架。本检测系统不会对原始的DNN模型进行任何修改, 所以不会影响原始模型的识别性能。此外, 本文仅使用干净样本作为FIAD的训练数据, 因此本系统属于不局限于特定对抗攻击的无监督检测方法。FIAD的整体流程为训练检测模型后对对抗样本进行检测, 将在下文详细介绍。

2.3.2 基于特征不变量的检测模型

为了叙述方便, 本文以 f_k 表示DNN的第 k 层, $f_k = \sigma(w_k \cdot x_k + b_k)$, 其中, σ 为激活函数, w_k 为模型权重, b_k 为偏置, x_k 为该层的输入, 则具有 n 层的DNN可记为 $F_n = f_n \circ f_{n-1} \circ f_{n-2} \circ \dots \circ f_1$, 其中最后一层 f_n 的激活函数一般为Softmax。

由于不同类别的干净数据在DNN模型中的神经元激活状态不同, 因此本文为每类数据都构建独立的特征不变量模型。对于类别标签为 i 的干净样本, 收集其在DNN模型中每层的激活值和LID, 以用作类 i 的检测模型的训练数据。

首先为所有类 i 的干净输入的每一层训练一个

以激活值为标准的激活值不变量模型, 这些模型表征了该类干净样本在DNN中的激活值分布, 记作 VAD_i^l (value activation distribution), $l \in [1, n]$, 其模型结构为:

$$\text{VAD}_i^l(x_i) = f_1(x_i) \circ f_2 \cdots \circ f_{l-2} \circ f_{l-1} \circ f_l \circ w_i^l \quad (6)$$

对于层 l 而言, $F_l(x_i)$ 即为干净样本 x_i 在该层经过激活函数转换后的激活值分布状态, 而 $\text{VAD}_i^l(x_i)$ 模型是判断基于层 l 的激活值预测模型的输入 x_i 是否为干净样本。因此, 学习 $\text{VAD}_i^l(x_i)$ 模型的权重 w_i^l 即解决优化问题:

$$\min_{x_i \in X_i} \sum J(f_1(x_i) \circ f_2 \cdots \circ f_{l-2} \circ f_{l-1} \circ f_l \circ w_i^l - 1) \quad (7)$$

其中, X_i 为所有类 i 的干净数据, $J(\cdot)$ 为损失函数。本文旨在最大限度地提高干净输入 x_i 在 VAD_i^l 中输出为1的概率。因此, 对于共有 n 层的神经网络模型, 类 i 的激活值不变量模型 $\text{VAD}_i = [\text{VAD}_i^1, \text{VAD}_i^2, \dots, \text{VAD}_i^n]$, 每个 VAD_i^l 均由式(6)和式(7)独立得到。

相应地, 用LID不变量模型 VLID_i (value local intrinsic dimensionality) 来表示属于类 i 的不同干净样本的每层激活值之间的局部空间维度, 以表征该类干净样本在神经网络中彼此之间的约束复杂程度, 其模型结构为:

$$\text{VLID}_i = [\text{LID}_i^1, \text{LID}_i^2, \dots, \text{LID}_i^n] \circ w_i \quad (8)$$

学习 VLID_i 模型的权重 w_i 即解决优化问题:

$$\min_{x_i \in X_i} \sum J([\text{LID}_i^1, \text{LID}_i^2, \dots, \text{LID}_i^n] \circ w_i - 1) \quad (9)$$

值得注意的是, VAD_i 和 VLID_i 在训练过程中需要固定原始模型的权重参数, 并重新训练类

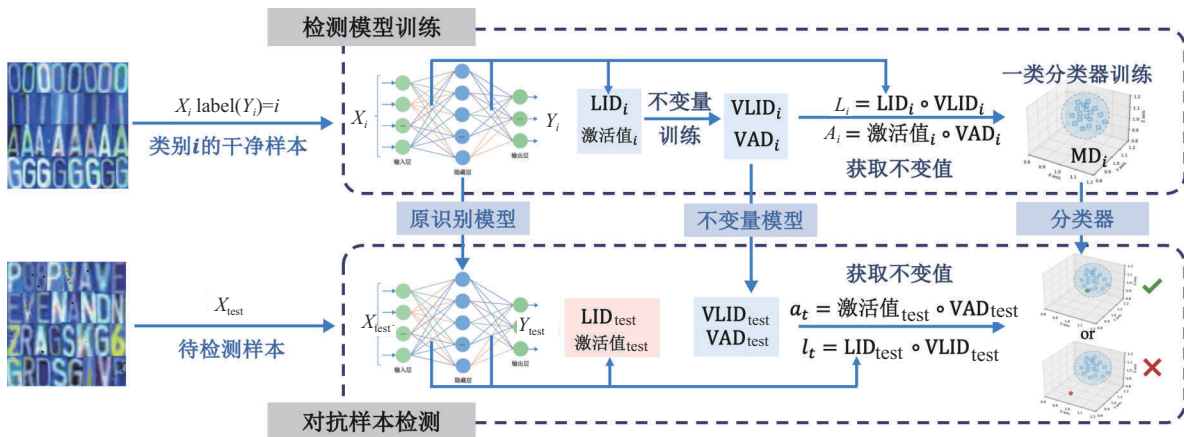


图4 FIAD系统框架

Figure 4 FIAD system framework

似神经网络全连接层的权重 w 。由于仅训练一层，因此通常只需消耗有限的时间即可完成训练。

此外，与同是利用神经网络不变性的NIC^[16]检测系统相比，本文使用 VAD_i 即可实现 JD (joint distribution) 和 VD (value distribution) 的检测效果，不再需要利用 JD 构建多个派生模型。因为 NIC 中 JD 会将相邻两层的预测结果相关联，以提供汇总两层激活值分布之间相关性的方法，即 JD 将模型较浅层的原始特征与相应的较后层的抽象特征相关联。对于 VAD_i ，其本身即描述了单一输入类别的激活值分布状态。以图2中神经网络为例，图5展示了对于A和B类的VD、JD及 VAD_i 的训练激活值分布状态。对于图2(d)中被误识为B的A类样本来说，其 Lay_1 和 Lay_2 的激活值状态均在 NIC 的 VD 模型的正常范围内，而 JD 却能汇总两层激活值间的关系，识别出此时 Lay_1 和 Lay_2 整体并不是正常分布，如图5(a)红色部分所示。而 FIAD 的 VAD_B 中并不存在此时 Lay_1 和 Lay_2 的分布状态如图5(b)所示。因此，按原始模型的标签分类，为每类单独训练一个 VAD_i 以代替 JD。同时，还添加 VLID_i 来进一步加强检测效果。

对于特定的类 i ，可得 VAD_i 和 VLID_i 两种特征不变量检测模型。以类 i 的干净数据 X_i 作为检测模型输入，以提取类 i 的干净样本的不变量特征 A_i^l 和 L_i 。本文的目标是以干净样本的不变量为标准，构建不需要获得有关攻击的先验知识的通用检测方法。因此，检测模型的训练问题实际上可视为一类分类 (one-class classification, OCC) 问题。在 OCC 中，训练样本都是良性的，而在测试过程中输入包括来自各种攻击的对抗样本。虽然 OCC 通常不如利用正负样本的技术准确，但使用多个不变量模型进行联合决策，可以有效缓解

OCC 的分类误差，从而确保足够高的准确率。

本文使用 OCC 中流行的分类算法——一类支持向量机 (one-class support vector machine, OSVM)^[36]，其专门设计用于处理只有单一类别数据的情况。OSVM 的基本思想是假设正负类之间的边界形状 (由不同的核函数表示)，进而计算描述该边界形状的参数。本文中类 i 的干净样本的不变量特征 $A_i^l (l \in [1, n])$ 和 LID_i 基本聚集在较小的子空间中，使用径向基核函数 (radial basis function, RBF) 可获得良好的准确率。对于类 i ，以 A_i^l 和 LID_i 作为良性输入 V_i ，训练以 OSVM 为基础的检测分类模型 MD_i 即解决优化问题：

$$\min_{w, \xi, \rho} \frac{1}{2} \|w\|^2 + \frac{1}{vn_i} \sum_{k=1}^{n_i} \xi_k - \rho \quad (10)$$

$$\text{s.t. } w \cdot \phi(v_i) \geq \rho - \xi_k, \xi_k \geq 0, \forall v_i \in V_i$$

其中， w 是权重向量， ξ_k 是松弛变量， ρ 为决策函数的偏移项， v 为控制支持向量的比例和决策函数的平滑度参数，训练样本的数量为 $|V_i| = |X_i| = n_i$ ，即类 i 的干净样本的总数。 $\phi(v_i)$ 是由 RBF 核定义的将输入数据映射到高维空间的函数。最终，在检测模型的训练阶段，每类干净数据将会训练得到自己独立的检测分类器 MD_i 。

2.3.3 对抗样本检测

在系统运行期间，对于待检测的样本 X_{test} ，将其输入原始识别模型可得预测输出 Y_{test} ，同时选取 Y_{test} 类对应的不变量模型 VAD_{test}^l 和 VLID_{test} 以获得该样本的检测特征。具体来说，本文收集 X_{test} 在原始模型中每层的激活值和 LID，并将其作为对应类不变量检测模型 VAD_{test}^l 和 VLID_{test} 的输入，得到相应的特征值 a_i^l 和 l_i 。以 $v_i = [a_i^1, \dots, a_i^n, l_i]$ 作为不变量模型的联合特征表示，输入对应类的检

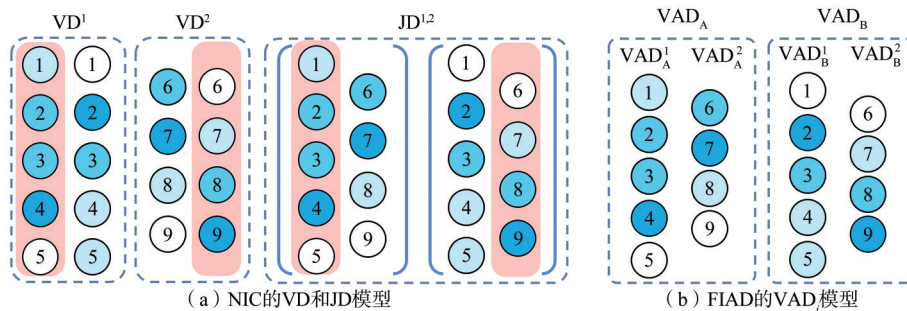


图5 NIC和FIAD不变量模型的训练激活值对比

Figure 5 Comparison of training activation values between NIC and FIAD invariant models

测分类器 MD_{test} 中, 最终分类器会对 v_i 进行联合决策以判断 X_{test} 是否为对抗样本。

本文取干净样本和对抗样本的 $VAD'(l \in [1, 2, 3])$ 的输出特征值得到三维散点图 (对抗样本与干净样本的不变量模型输出示例如图6所示), 可以发现干净样本的散点聚集在较小的范围内, 具有特征不变性。然而, 对抗样本的散点分散在三维空间中, 说明它们的不变量存在特征异常。此时, 检测分类器 (MD) 的边界如图6中绿色阴影所示, 其通过判断样本的特征值点是否在正常的边界范围内, 从而决定样本是否为对抗样本。图6表明了检测分类器 (MD) 具有良好的检测准确率。

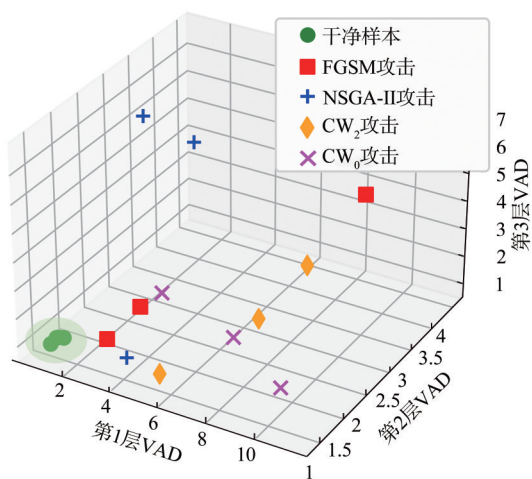


图6 对抗样本与干净样本的不变量模型输出示例
Figure 6 Invariant model output examples for adversarial and clean samples

2.4 FIAD在车牌识别场景中的适用性

本文提出的FIAD能很好地应用于车牌识别场景, 这主要是因为车牌识别存在识别种类有限以及背景颜色统一的特点。

1) 识别种类有限。车牌识别场景仅需识别字母、数字及汉字, 种类较为有限, FIAD的分类检测器训练较为容易。但在处理复杂场景如大规模人脸识别时分类检测器的训练难度将大幅增加。

2) 背景颜色统一。车牌主要背景色为蓝色或绿色, 这种单一背景色减少了对识别主体的干扰, 使FIAD能更容易从神经元的激活状态中获取和学习特征的不变性。

本文在HyperLPR和EasyPR这两个中国车牌识别系统上的实验, 进一步有效验证了FIAD在

车牌识别场景中的高可信性和有效性。

3 车牌识别场景的实验及分析

3.1 车牌识别系统

本文选择了国内应用最为广泛的HyperLPR^[5]和EasyPR^[6]两个针对中国车牌的开源识别模型开展实验, 这两个模型都可识别多种场景下的车牌。

在实际场景中, 车牌识别系统的识别流程一般为如图7所示的4个步骤。本文利用HyperLPR的分割模型将原始图像中的车牌进行分割, 最终得到车牌字符数据集, 并以此为基础创建对抗样本攻击HyperLPR和EasyPR的车牌字符识别模块。



图7 车牌识别系统的识别流程
Figure 7 The recognition process of license plate recognition system

3.2 CCPD

CCPD (Chinese city parking dataset)^[17] 是目前为止最大的开源中国车牌数据集。经过预处理, 本实验的车牌数据集由约2.5万张车牌构成, 通过车牌字符分割后得到约17万张的车牌字符数据集。本文将整个字符数据集随机划分为训练集 (80%用于训练检测模型) 和测试集 (20%用于生成对抗样本以测试检测模型的最终性能)。

3.3 实验准备

3.3.1 对抗样本攻击

本文对主流的对抗样本攻击针对车牌字符数据集和两个车牌识别系统进行了实现, 其中基于梯度的对抗攻击有FGSM^[7]、BIM^[19]、CW^[18] (包括基于 L_0 、 L_2 和 L_∞ 的3种形式) 和DeepFool^[20]。基于内容的对抗攻击有DeepXplore^[21] (包括Light、Dirt和RP的3种形式, 图8中横坐标省略DeepXplore) 和基于NSGA-II的物理攻击^[23]。

此外, 还选择了专门针对车牌识别系统的斑点规避攻击 (spot evasion attack, SEA)^[37], 其在车牌的特定区域添加预定的扰动, 以模拟某种自

然形成的斑点，如泥浆等。

3.3.2 对比检测方案

本文将FIAD与目前先进的4种检测技术进行了对比，包括基于统计测量的LID^[14]、基于辅助模型的NSS（natural scene statistic）^[25]、基于预测不一致的特征压缩（feature squeezing, FS）法^[15-30]和基于网络不变性的NIC^[16]。在FS法中，本文选择位深度（bit depth, BD）和自适应噪声减少（adaptive noise reduction, ANR）两种压缩方式。

3.4 不同方法的检测效果对比

图8和表1为不同攻击下不同检测器的假阳率和平均检测准确率，其中检测准确率为正确检测出的对抗样本数占有对抗样本的比例，即二分类的召回率。这些攻击方法均是基于原始识别模型展开，不考虑本文的检测方案，因此属于1.4节威胁模型中的情况1。结果表明，FIAD能够以较低的假阳率（HyperLPR为7.69%，EasyPR为8.86%）检测所有基于梯度和内容攻击生成的对抗样本，且平均检测准确率高达99%，而其他检测器则有较高的假阳率或较低的准确率。例如，对于如图8所示的EasyPR模型，NIC对所有攻击的检测准确率均略低于FIAD。因此，相较其他检测

器，本文的FIAD具有较高的可信性和有效性。

表1 不同检测方法效果对比
Table 1 Comparison of the effectiveness of different detection methods

检测方法	识别模型	假阳率	平均检测准确率
FIAD	HyperLPR	7.69%	99.98%
NIC		10.11%	99.93%
LID		17.75%	90.08%
NSS		11.25%	96.12%
FS(BD)		12.89%	92.07%
FS(ANR)		10.75%	93.55%
FIAD	EasyPR	8.86%	99.74%
NIC		11.15%	99.14%
LID		18.85%	89.74%
NSS		11.66%	94.79%
FS(BD)		14.61%	89.09%
FS(ANR)		12.23%	90.64%

此外，本文将FIAD与同是基于神经网络不变性思想的NIC进行了详细的对比，如表2所示，其中划分准确率为分类正确的样本数占总样本的比例。相较NIC，FIAD在不变量模型和分类器的训练时间上明显更短。这是因为NIC除了

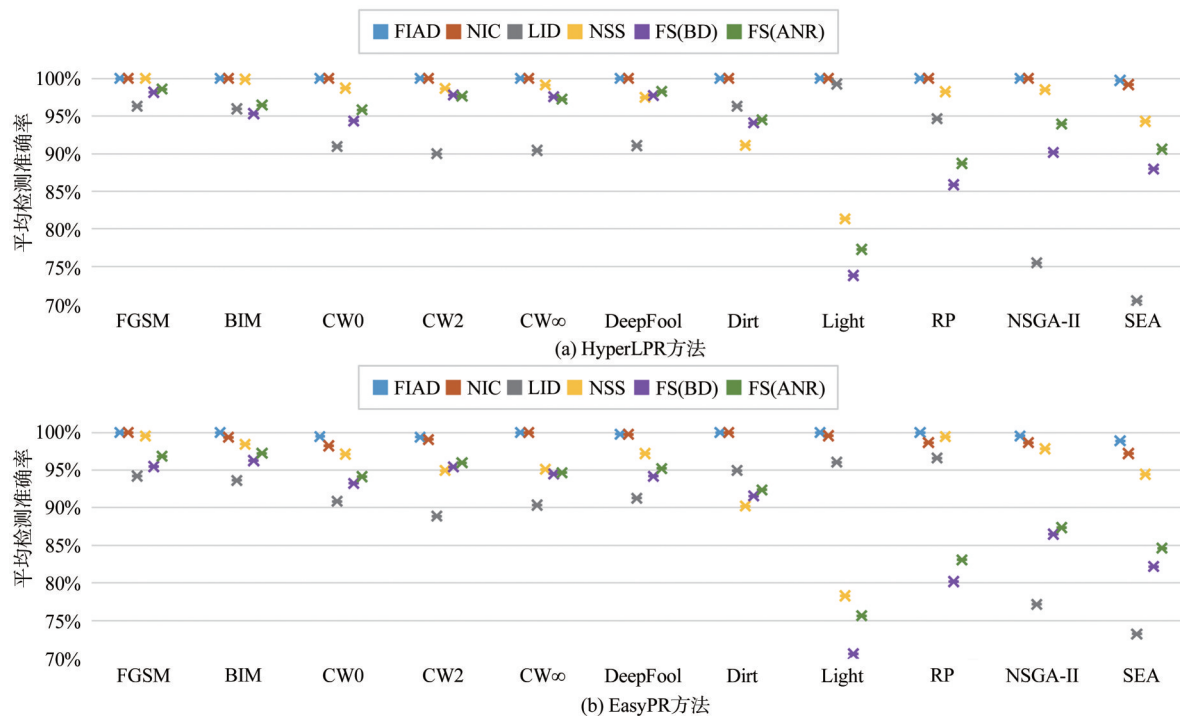


图8 不同检测方法平均检测准确率对比
Figure 8 Comparison of average detection accuracy of different detection methods

表2 基于网络不变量的检测方案FIAD与NIC对比
Table 2 Comparison between FIAD and NIC based on network invariants

检测方法	识别模型	平均时间			平均准确率		
		分类器训练/s	训练不变量模型/h	检测每个样本/s	假阳率	检测准确率	划分准确率
NIC	HyperLPR	44.32	50.25	0.61	10.11%	99.93%	94.91%
FIAD		2.34	35.80	0.59	7.69%	99.98%	96.15%
NIC	EasyPR	26.98	14.71	0.46	11.15%	99.14%	94.00%
FIAD		1.87	10.68	0.44	8.86%	99.74%	95.44%

需要训练VAD外，还需要构建多个派生子模型以联合训练JD，FIAD仅用VAD_i即可实现VAD和JD的检测效果，训练时间大量减少。

在准确率上，FIAD略高于NIC，这是因为FIAD按标签划分后针对每类数据单独训练不变量模型，而NIC则是针对所有的干净数据构建不变量模型，可能存在训练数据量大、特征复杂等问题，从而影响了准确率。

3.5 自适应白盒攻击评估

FIAD本身也属于机器学习算法，同样也会受到对抗样本的攻击^[34]。假设攻击者对本文的检测方案有全面的了解，即1.4节威胁模型中的情况2，则此时的自适应白盒攻击可视为针对FIAD检测的对抗攻击，即攻击者需要不断优化能被FIAD成功检测的初始对抗样本 x_{adv} ，以使优化后的样本 x'_{adv} 能够成功躲避FIAD的检测。本文对FIAD的对抗样本生成过程做了简化，即利用梯度上升思想生成针对VAD模型的对抗性样本。因此， x'_{adv} 生成的优化过程可简化为：

$$x_{temp} = x_k + \alpha \cdot \nabla_{x_k} \left(\sum_{l=1}^n J(f_1(x_k) \circ \dots \circ f_l \circ w'_l - 1) \right),$$

$$x_{k+1} = \min(\max(x_{temp}, -1), 1) \quad (11)$$

其中， $k \in [0, T-1]$ ， T 为优化次数，实验中为800。 $x_0 = x_{adv}$ 为待优化初始对抗样本，且 x_{adv} 的误识标签为 i ， $x_T = x'_{adv}$ 。对于VAD $^l(l \in [1, n])$ 共 n 个模型，本文以损失函数联合相加的方式求优化梯度。 $x_{k+1} \in [-1, 1]$ 以控制样本在有效范围内。

自适应白盒攻击生成的对抗样本如图9所示，第一行展示了标签为“R”的原始图像，第二行为对抗攻击生成的针对原车牌识别系统（HyperLPR）的对抗样本，它们均被识别模型误识为“P”，但却能被FIAD成功检测。以第二行为待优化的初始对抗样本 x_{adv} ，利用式(11)优化后得到第三行的图像 x'_{adv} ，且此时的 x'_{adv} 已经能成功躲避

FIAD针对标签“P”的不变量检测。观察发现，第三行的图像 x'_{adv} 相对原始图像变化明显，平均 L_2 失真相对 x_{adv} 增大了1.85倍，且此时部分标签又回归到原始正确的标签。因此，FIAD显著提高了攻击者构建对抗样本的难度。

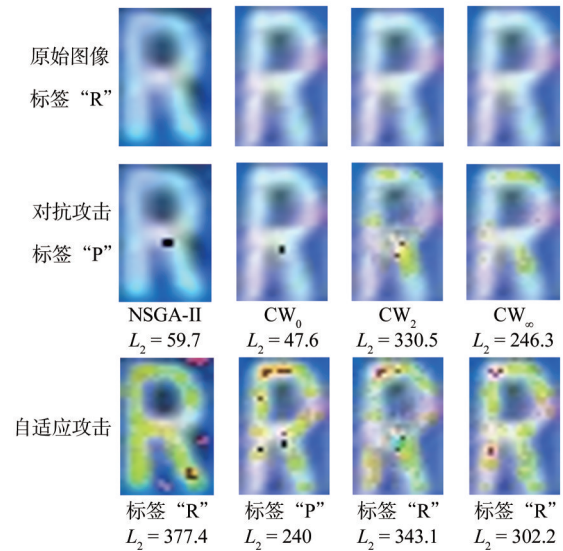


图9 自适应白盒攻击生成的对抗样本
Figure 9 Adversarial samples generated by adaptive white-box attacks

4 应用场景扩展

基于特征不变量思想的FIAD检测方案除了能和车牌识别场景完美适配，也可扩展到其他对抗样本检测的应用场景中。本文以交通标志识别场景为例进一步说明FIAD的有效性。选取中国交通标志检测数据集（Chinese traffic sign detection benchmark, CCTSDB）^[38]，并以文献[39]中的学生网络作为原始识别模型，其识别准确率可高达99.61%。不同攻击在CCTSDB上的对抗样本示例如图10所示。实验表明，FIAD在CCTSDB上的对抗检测假阳率仅为4.7%。FIAD在CCTSDB上的对抗样本检测准确率如表3所示，FIAD在交通标志识别场景中也适用。



图10 不同攻击在CCTSDB上的对抗样本
Figure 10 Adversarial samples of different attacks on the CCTSDB

表3 FIAD在CCTSDB上的对抗样本检测准确率
Table 3 Detection accuracy of FIAD on CCTSDB

FGSM	BIM	DeepXplore			NSGA-II
		Dirt	Light	RP	
99.1%	96.9%	99.4%	97.9%	98.2%	97.6%

5 结束语

本文针对真实场景下车牌识别系统的对抗样本攻击，通过分析神经网络在干净样本上训练的内在变化和干净样本之间的维度复杂度，提出了基于神经网络不变量和局部固有维数不变量的特征不变量无监督对抗样本检测系统FIAD。在车牌识别场景下，FIAD的无监督性也使本文无须考虑复杂自然环境对检测系统带来的各种挑战。

虽然FIAD相比其他的检测技术具有较高的检测准确率和较低的假阳率，但仍有一定程度的假阳率，这可能与检测模型训练时过拟合以及数据本身的迷惑性有关。因此，本系统仍有一定的改进空间。此外，本文的特征不变量思想可以扩展到其他对抗样本检测的应用场景中，如交通标志识别，这为对抗检测技术的研究提供了新的思路。

参考文献：

- [1] ZHAO H S, JIA J Y, KOLTUN V. Exploring self-attention for image recognition[C]//Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2020: 10073-10082.
- [2] LI J Y. Recent advances in end-to-end automatic speech recognition[EB]. arXiv Preprint, arXiv: 2111.01690, 2022.
- [3] LUO W H, XING J L, MILAN A, et al. Multiple object tracking: a literature review[J]. Artificial Intelligence, 2021(293): 103448.
- [4] GEHR T, MIRMAN M, DRACHSLER-COHEN D, et al. AI2: safety and robustness certification of neural networks with abstract interpretation[C]//Proceedings of the 2018 IEEE Symposium on Security and Privacy (SP). Piscataway: IEEE Press, 2018: 3-18.
- [5] YAN J, YU J, XIAOXIAO. HyperLPR3 - high performance license plate recognition framework[EB]. 2023.
- [6] LIURUOZE, MICOOZ, JSXYHELU, et al. An easy, flexible, and accurate plate recognition project for Chinese licenses in unconstrained situations[EB]. 2023.

- [7] GOODFELLOW L J, SHLENS J, SZEGEDY C. Explaining and harnessing adversarial examples[EB]. arXiv Preprint arXiv: 1412.6572, 2014.
- [8] ALDAHDOOH A, HAMIDOUCH W, FEZZA S A, et al. Adversarial example detection for DNN models: a review and experimental comparison[J]. Artificial Intelligence Review, 2022, 55(6): 4403-4462.
- [9] MADRY A, MAKELOV A, SCHMIDT L, et al. Towards deep learning models resistant to adversarial attacks[EB]. arXiv Preprint arXiv: 1706.06083, 2017.
- [10] BORKAR T, HEIDE F, KARAM L. Defending against universal attacks through selective feature regeneration[C]//Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2020: 706-716.
- [11] LIAO F Z, LIANG M, DONG Y P, et al. Defense against adversarial attacks using high-level representation guided denoiser[C]//Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018: 1778-1787.
- [12] 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 Security. New York: ACM, 2017: 506-519.
- [13] PAPERNOT N, MCDANIEL P, JHA S, et al. The limitations of deep learning in adversarial settings[C]//Proceedings of the 2016 IEEE European Symposium on Security and Privacy (EuroS&P). Piscataway: IEEE Press, 2016: 372-387.
- [14] MA X J, LI B, WANG Y S, et al. Characterizing adversarial subspaces using local intrinsic dimensionality[EB]. arXiv Preprint arXiv: 1801.02613, 2018.
- [15] XU W L, EVANS D, QI Y J. Feature squeezing: detecting adversarial examples in deep neural networks[C]//Proceedings 2018 Network and Distributed System Security Symposium. Reston, VA: Internet Society, 2018.
- [16] MA S Q, LIU Y Q, TAO G H, et al. NIC: detecting adversarial samples with neural network invariant checking[C]//Proceedings 2019 Network and Distributed System Security Symposium. Reston, VA: Internet Society, 2019.
- [17] XU Z B, YANG W, MENG A J, et al. Towards end-to-end license plate detection and recognition: a large dataset and baseline[C]//Proceedings of the Computer Vision-ECCV 2018: 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XIII. New York: ACM, 2018: 261-277.
- [18] CARLINI N, WAGNER D. Towards evaluating the robustness of neural networks[C]//Proceedings of the 2017 IEEE Symposium on Security and Privacy (SP). Piscataway: IEEE Press, 2017: 39-57.
- [19] KURAKIN A, GOODFELLOW L, BENGIO S. Adversarial examples in the physical world[M]. Artificial Intelligence Safety and Security: Chapman and Hall/CRC Press, 2018.
- [20] MOOSAVI-DEZFOOLI S M, FAWZI A, FROSSARD P. DeepFool: a simple and accurate method to fool deep neural networks[C]//Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2016: 2574-2582.
- [21] PEI K X, CAO Y Z, YANG J F, et al. DeepXplore: automated whitebox testing of deep learning systems[C]//Proceedings of the 26th Symposium on Operating Systems Principles. New York: ACM, 2017: 1-18.

- [22] LIU Y Q, MA S Q, AAFER Y, et al. Trojaning attack on neural networks[C]//Proceedings 2018 Network and Distributed System Security Symposium. Reston, VA: Internet Society, 2018.
- [23] MAO X F, CHEN Y F, WANG S H, et al. Composite adversarial attacks[C]//Proceedings of the AAAI Conference on Artificial Intelligence, 2021, 35(10): 8884-8892.
- [24] MUSTAFA A, KHAN S H, HAYAT M, et al. Image super-resolution as a defense against adversarial attacks[J]. IEEE Transactions on Image Processing: A Publication of the IEEE Signal Processing Society, 2019: 1711-1724.
- [25] KHERCHOUCHE A, FEZZA S A, HAMIDOUCE W, et al. Detection of adversarial examples in deep neural networks with natural scene statistics[C]//Proceedings of the 2020 International Joint Conference on Neural Networks (IJCNN). Piscataway: IEEE Press, 2020: 1-7.
- [26] ZUO F, ZENG Q. Exploiting the sensitivity of L2 adversarial examples to erase-and-restore[C]//Proceedings of the 2021 ACM Asia Conference on Computer and Communications Security. New York: ACM, 2021: 40-51.
- [27] ALDAHDOOH A, HAMIDOUCE W, DÉFORGES O. Revisiting model's uncertainty and confidences for adversarial example detection[J]. Applied Intelligence, 2023, 53(1): 509-531.
- [28] LI X, LI F X. Adversarial examples detection in deep networks with convolutional filter statistics[C]//Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV). Piscataway: IEEE Press, 2017: 5775-5783.
- [29] COHEN G, SAPIRO G, GIRYES R. Detecting adversarial samples using influence functions and nearest neighbors[C]//Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2020: 14441-14450.
- [30] LIANG B, LI H C, SU M Q, et al. Detecting adversarial image examples in deep neural networks with adaptive noise reduction[J]. IEEE Transactions on Dependable and Secure Computing, 2021, 18(1): 72-85.
- [31] ENISER H F, CHRISTAKIS M, WÜSTHOLZ V. RAID: randomized adversarial-input detection for neural networks[EB]. arXiv Preprint arXiv: 2002.02776, 2020.
- [32] MENG D Y, CHEN H. MagNet: a two-pronged defense against adversarial examples[C]//Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security. New York: ACM, 2017: 135-147.
- [33] FREITAS S, CHEN S T, WANG Z J, et al. UnMask: adversarial detection and defense through robust feature alignment[C]//Proceedings of the 2020 IEEE International Conference on Big Data. Piscataway: IEEE Press, 2020: 1081-1088.
- [34] BIGGIO B, ROLI F. Wild patterns: ten years after the rise of adversarial machine learning[J]. Pattern Recognition, 2018, 84(C): 317-331.
- [35] AMSALEG L, CHELLY O, FURON T, et al. Estimating local intrinsic dimensionality[C]//Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2015: 29-38.
- [36] ALAM S, SONBHADRA S K, AGARWAL S, et al. One-class support vector classifiers: a survey[J]. Knowledge-Based Systems, 2020(196): 105754.
- [37] QIAN Y G, MA D F, WANG B, et al. Spot evasion attacks: adversarial examples for license plate recognition systems with convolutional neural networks[J]. Computers & Security, 2020, 95: 101826.
- [38] ZHANG J, XIE Z, SUN J, et al. CSUST Chinese traffic sign detection benchmark[EB]. 2024.
- [39] ZHANG J M, WANG W, LU C Q, et al. Lightweight deep network for traffic sign classification[J]. Annals of Telecommunications, 2020, 75(7): 369-379.

[作者简介]



朱孝羽（2000-），女，湖北宜昌人，上海交通大学硕士生，主要研究方向为人工智能安全、联邦学习。



唐鹏（1992-），男，江西抚州人，上海交通大学助理研究员，主要研究方向为人工智能安全、隐私保护。



张浩臣（1999-），男，江西九江人，上海交通大学硕士生，主要研究方向为人工智能安全。



邱卫东（1973-），男，江西修水人，上海交通大学教授、博士生导师，主要研究方向为计算机取证、密码分析、人工智能安全、大数据隐私保护。



黄征（1975-），男，四川南充人，博士，上海交通大学副教授，主要研究方向为隐私保护、计算机视觉、人工智能安全。