基于局部邻域滤波的对抗攻击检测方法

刘 朝[1,2], 朱莉芳[1,2], 接 标[1,2], 丁新涛[1,2]

[1] 安徽师范大学计算机与信息学院 芜湖 中国 241002
[2] 网络与信息安全安徽省重点实验室 芜湖 中国 241002

摘要 目前, 卷积神经网络在语音识别、图像分类、自然语言处理、语义分割等方面都取得了良好的应用成果, 是计算机应用研究最广泛的技术之一。但研究人员发现当向输入中加入特定的微小扰动时, 卷积神经网络(CNN)模型容易产生错误的预测结果, 这类含有微小扰动的图像被称为对抗样本, CNN 模型易受对抗样本的攻击。对抗样本的出现可能会对安全敏感的领域带来潜在的应用威胁。已有较多的防御方法被提出, 其中许多方法对特定攻击方法具有较好的防御效果, 但由于实际应用中无法知晓攻击者采用的攻击方式, 因此提出不依赖攻击方法的通用防御策略是一个值得研究的问题。为有效地防御各类对抗攻击, 本文提出了基于局部邻域滤波的对抗攻击检测方法。首先, 通过像素间的相关性对图像进行 RGB 空间切割。其次将相似的图像块组成立方体。然后, 基于立方体中邻域的局部滤波进行去噪, 即: 通过邻域立方体的 3 个块得到邻域数据的 3 维标准差, 用于 Wiener 滤波。再将滤波后的块组映射回 RGB 彩色空间。最后, 将未知样本和它的滤波样本分别作为输入, 对模型的分类进行一致性检验, 如果模型对他们的分类不相同, 则该未知样本为对抗样本, 否则为良性样本。实验表明本文检测方法在不同模型中对多种攻击具备防御效果, 识别了对抗样本的输入, 且在 mini-ImageNet 数据集上针对 C&W、DFool、PGD、TPGD、FGSM、BIM、RFGSM、MI-FGSM 以及 FFGSM 攻击的最优检测结果分别达到 0.938、0.893、0.928、0.922、0.866、0.840、0.879、0.889 以及 0.871, 结果表明本文方法在对抗攻击上具有鲁棒性和有效性。

关键词 卷积神经网络; 对抗攻击; 局部邻域滤波; 对抗检测
中图法分类号 TP309.2 DOI 号 10.19363/J.cnki.cn10-1380/tn.2023.11.07

Adversarial Attack Detection Method Based on Local Neighborhood Filtering

LIU Chao[1,2], ZHU Lifang[1,2], JIE Biao[1,2], DING Xintao[1,2]

[1] School of Computer and Information, Anhui Normal University, Wuhu 241002, China
[2]Anhui Provincial Key Laboratory of Network and Information Security, Wuhu 241002, China

Abstract Currently, convolutional neural networks (CNN) are greatly attributed in speech recognition, image classification, natural language processing, and semantic segmentation. They are widely studied in computer applications. However, CNN models are vulnerable to adversarial examples that have been crafted specifically to fool a system while being imperceptible to humans. Such images containing small perturbations are called adversarial examples. Adversarial examples may pose a potential threat to security-sensitive applications. In this study, we focus on adversarial defense on CNN models. Since attack method is usually unknown for server in practical applications, proposing a general defense method that does not depend on attack method is an interesting topic. In order to effectively defend against various types of adversarial attacks, this paper proposes an adversarial attack detection method based on local neighborhood filtering. Firstly, the input image is divided in similar regions using inter-pixel correlation after the pixel values are projected in RGB space. Secondly, every similar region is rearranged to a block based on pixel intensity, and the image blocks are formed into a cube. After obtaining the 3-dimensional standard deviation of the neighborhood data in the cube, Wiener filtering is then performed based on the local filtering in the neighborhood of the cube. After that, the filtered block set is converted into RGB space. Finally, the input and its filtered example are respectively fed to CNN model for classification. If the model classifies them to different classes, the input example is taken as an adversarial example. Otherwise, it is discriminated as a benign example. The comparison experiments show that our proposed method is effective against adversarial attacks on different models. The detection rates on the mini-ImageNet dataset against C&W, DFool, PGD, TPGD, FGSM, BIM, RFGSM, MI-FGSM, and FFGSM attacks are 0.938, 0.893, 0.928, 0.922, 0.866, 0.840, 0.879, 0.889 and 0.871, respectively. The results show our method is robust and effective against adversarial attacks.

Key words convolutional neural networks; adversarial attack; local neighborhood filtering; adversarial detection

通讯作者: 丁新涛, 博士, 副教授, Email: dincent@ahnu.edu.cn。
本课题得到安徽省自然科学基金面上项目(No. 1808085MF171), 以及国家自然科学基金面上项目(No. 61976006)的资助。
收稿日期: 2022-02-25; 修改日期: 2022-08-09; 定稿日期: 2023-08-26

1 引言

卷积神经网络(Convolutional Neural Networks, CNN)是深度学习的代表性模型之一, 在语义分割[1]、自动驾驶[2]、动作跟踪[3]、语音识别[4]等方面的应用中取得了巨大的成功。然而, 2014 年 Szegedy 等人发现卷积神经网络容易受到精心制作的微小扰动的影响。在图像分类中对抗攻击的定义: 给定一张正确分类的图像, 攻击者向其添加精心制作的微小扰动得到对抗样本, 使得神经网络模型分类错误并以较高置信度误判的恶意攻击。在自动驾驶、语音识别、文字识别等应用领域, 对抗样本可成功实现较高概率的攻击, 有效防御对抗攻击对于进一步推广神经网络模型应用具有重要意义。

自发现 CNN 模型易遭受对抗样本的攻击之后, 研究人员进行了大量的针对 CNN 模型的对抗攻击研究工作, 并提出了许多对抗攻击方法。在这些方法中, 既有利用梯度信息的 FGSM[5]和 BIM[6]攻击, 也有基于优化的 C&W[7]方法, 除此之外, 还有基于迁移的攻击[8]、基于 GAN 的攻击[9]等。由于对抗攻击通过向良性样本添加“有目的”的噪声削弱图像的特征信息, 以欺骗 CNN 模型。对抗攻击总是在输入添加扰动, 因此, 对输入进行特定的变换、重建能够防御对抗攻击。

本文的主要动机在于研究通过图像重建是否可以破坏对抗扰动, 进而达到对抗防御的效果。基于此, 本文提出一种基于局部图像相似块分组的滤波降噪方法, 用于对抗攻击的检测。其主要思想是通过 RGB 空间切割得到相似块立方体, 并在局部立方体上设计一个三维傅里叶滤波进行图像重建, 通过输入变换进行对抗攻击的检测。

图 1 显示了本文方法的防御示例, 从左到右, 依次为原图(良性样本)、对抗样本、对抗样本中的噪声(对抗扰动)、滤波样本(经本文方法处理后的样本)和滤波样本的噪声(滤波样本和良性样本的差异), 为了展示效果, 本文特意可视化出对抗样本和滤波样本中的噪声, 从而体现本文方法的有效性。“原图”在 ResNet-50 模型中正确识别为 komondor, 在 FFGSM 攻击后, 对抗样本被误标为 broom。被攻击的样本中存在大量扰动, 相比于“原图”, 其特征轮廓模糊, 关键信息被干扰, 这是导致识别错误的原因之一。本文利用局部邻域滤波对对抗扰动进行破坏, 以抵御对抗攻击。虽然重建的滤波图像不如原图清晰(如图 1 中第四列所示), 但是对比图 1 第五列与第三列可知, 本文方法重建的图像破坏了对抗噪声, 如图 1 的第五列所示, 滤波样本中的噪声包含了物体的轮廓信息, 从而使得本文方法具有一定的防御效果。

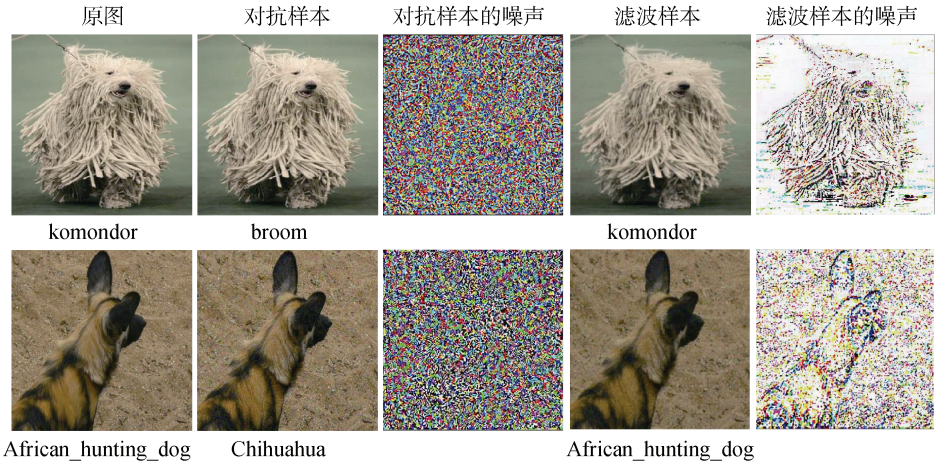

图 1 本文方法的防御示例

Figure 1 Example of defense of this method

本文的主要贡献如下:

1) 提出一种通过滤波将对抗扰动进行破坏, 进而进行对抗攻击检测的方法。通过图像空域分块, 在小块图像上进行频域滤波重建, 针对不同灰阶上的噪声生成不同参数进行滤波, 这能够破坏对抗噪声, 从而产生防御效果。

2) 本文设计了一种局部区域的三维傅里叶图像滤波算法, 在局部邻域内生成不同的参数进行滤波重建。

本文其余部分组织如下, 第 2 节介绍了对抗攻击与防御的相关工作; 第 3 节说明了基于局部邻域滤波的检测方法; 第 4 节验证了本文检测方法的性

能, 最后对文章进行总结。

2 相关工作

2.1 对抗攻击

对抗攻击指的是对目标模型的原输入添加微小扰动以生成对抗样本来愚弄目标模型的过程。一般来说, 可分为两种类型的攻击: 黑盒攻击和白盒攻击。白盒攻击需知晓目标模型的先验知识, 而黑盒攻击则无需知晓。

2.1.1 白盒攻击

已有多种白盒攻击方法被提出, 我们选取一些使用广泛的攻击作为本文的基础。

FGSM[5]是一个单步攻击过程。在白盒环境下, 求出模型对输入 x 的梯度, 用符号函数得到灰阶攻击方向, 接着乘以一个步长, 从而获得对抗扰动, 然后将其加在原来的输入 x 上, 得到 FGSM 攻击下的样本 $\hat{x}$。

BIM[6]是 FGSM 的迭代版本。每次迭代后, 生成的图像被裁剪到原始图像的一个 ϵL_∞ 邻域内, 找到对抗样本时, 停止此过程。

DFool (DeepFool [10])是一个无目标迭代攻击。该方法认为分类器是一个线性决策边界, 攻击时, 需要找到穿过该边界所需的最小扰动。

Carlini&Wagner (C&W[7]**)**是最强的攻击之一。C&W 更新损失函数, 使得 p – 范数下的损失 L_p 和一个自定义可微损失函数同时最小化, 且该损失函数使用分类器的非标准化输出。

PGD[11]也是 FGSM 的迭代版本。每次迭代都会将扰动裁剪到规定范围内。

MI-FGSM[12]是基于动量(Momentum)的 FGSM 的迭代版本。该方法在迭代过程中动态地更新方向, 排除不合适的 local maxima(局部最大值), 产生可移性高的对抗样本。

2.1.2 黑盒攻击

黑盒攻击是在没有目标模型信息的情况下进行, 对抗扰动通常只根据输出标签和置信度来生成。目前的主要实现方法有基于迁移的攻击[13]和基于访问的攻击。基于迁移的方法在目标模型的类似模型上生成对抗样本, 并借助对抗样本的迁移性来攻击目标模型。基于访问的攻击则依赖于目标模型的类别置信度。根据访问过程中所获信息的特点, 基于访问的攻击可细分为基于分数的攻击[14]和基于决策的攻击[15]。基于决策的攻击首先获取数值较大的初始扰动, 并以此为基础在决策边界附近找出幅度更小的扰动来生成高质量的对抗样本。而基于分数的攻击需获取一个连续的预测分数, 这在许多实际场景中不太适用。

2.2 对抗防御

根据防御效果, 对抗防御的研究主要分为检测防御和重识别防御。检测防御方法对输入样本进行二分类, 判断其是否为对抗样本; 重识别防御可通过强化 CNN 模型来还原对抗样本的特征信息, 从而对其进行正确地分类。本文主要研究检测防御方法。

2.2.1 检测防御方法

现有的检测防御手段主要包括: (1) 对抗性检测网络方法。Metzen 等[16]研究出对抗性检测网络(AND, Adversary Detector Network), 其使用二元检测器网络来扩充预训练神经网络, 从而区分对抗样本和良性样本。AND 方法可以有效检测出 BIM、DeepFool 与 FGSM 攻击。(2) 基于度量的方法。严飞等[17]提出了一种基于边界值不变量的对抗攻击检测方法, 该方法通过拟合分布来寻找深度网络模型中的不变量, 从而检测出对抗样本。(3) 特征压缩。Xu 等[18]认为输入特征的高维度导致了较大的攻击空间, 基于此, 他们提出基于特征压缩的检测方法。该方法通过压缩特征去除不必要的数据信息, 并比较非压缩与压缩之后的预测结果, 以区分对抗样本和良性样本。(4) 逆交叉熵检测方法。Pang 等[19]提出使用新目标函数进行反向检测的逆交叉熵检测方法(RCE, Reverse Cross-Entropy), 此方法通过训练额外的神经网络以区分良性样本和对抗样本。相比于使用标准交叉熵作为目标函数的检测方法, RCE 不仅进行了对抗性检测, 也在总体上增强了模型的鲁棒性。(5) 基于自编码器的方法。MagNet 等[20]在原数据集上训练自编码器来代替特征压缩中的图像处理, 并通过比较自编码器处理后的图像与原输入图像的概率向量来检测对抗样本。

2.2.2 重识别防御方法

重识别防御方法主要包含: (1) 数据扩充。此类方法将生成的对抗样本加入训练集进行再训练, 从而提升模型的鲁棒性[21]。(2) 正则化方法。该方法主要使用“压缩”技术来减少深度网络模型的规模, 从而降低网络梯度的大小, 提高防御微小扰动的能力[22]。(3) 基于扰动优化的方法。通过分析数据、设计模型等找到影响识别结果的扰动信息, 对其进行优化操作, 强化样本中的特征信息, 实现对对抗样本的防御。陈晋音等[23]提出了一种基于通用逆扰动的对抗样本防御方法, 该方法通过学习原始数据集中的类相关主要特征, 生成通用逆扰动(Universal

Inverse Perturbation, UIP), 且 UIP 对攻击方法和样本数据都具有通用性, 即一个 UIP 可以对多种攻击下的所有对抗样本进行防御。(4) 基于数据预处理的方法。该方法利用了对抗样本空间的不稳定性, 通过对原数据进行预处理来消除对抗扰动的影响, 例如去噪特征映射方法[24]等。(5) 基于梯度消失和梯度爆炸的方法。该防御措施主要是对每一层的偏导数进行累积, 这导致梯度过大或过小, 使得攻击者无法准确衡量用于生成对抗样本的梯度信息。例如 PixelDefend[25]和 Defense-GAN[26]均在神经网络分类器前增添了一个生成网络, 使分类模型的网络结构变得很深, 有利于将对抗样本转化为良性样本。

2.2.3 滤波降噪

滤波降噪机制已被证明在对抗防御中是有效的, 一些学者在网络模型输出分类结果前, 设计去噪模块, 消除扰动带来的影响。Liang 等[27]使用自适应去噪方法来减少对抗扰动的影响, 该方法首先以适当的间隔对样本进行量化, 该间隔是通过样本的熵来确定的。同时, 还使用熵来确定是否对量化样本进行平滑处理, 仅当熵大于阈值时, 才会通过空间平滑滤波器对其进行平滑。Xu 等[17]探索了两种压缩图像特征的方法: 减少图像中每个像素的颜色深度和使用空间平滑滤波器来减少各个像素之间的差异。虽然该方法可以减缓对抗扰动的影响, 但却会使良性样本的识别率降低。Wu 等[28]利用普通噪声和维纳滤波来抑制对抗干扰, 他们的方法包括两个操作: 先将自然噪声添加到对抗样本中, 然后再使用自适应维纳滤波对图像进行去噪。Xie 等[22]在卷积网络的中间层加入特征去噪模块, 用滤波对特征图进行降噪, 以增强模型的鲁棒性。然后再联合该去噪模块以端到端的方式进行对抗训练, 从而抑制特征图(Feature Map)上多余的噪声, 使响应集中在主要内容上。通过结合图像重建和去噪是减轻扰动的另一方法。Gupta 等[29]提出 CIIDefence, 该方法选择对当前分类结果影响最大的图像区域重建, 即选择类激活映射较高的几个区域进行重建。然后, 对剩下的未重建的区域进行基于小波变换的图像去噪, CIIDefence 不需要对分类器进行再训练或修改。

以上几种方法, 都采用降噪措施去消除扰动带来的影响, 即对特征图降噪、分区重建降噪或者量化后降噪等。但他们的方法是存在一些问题, 如很难准确地度量特征图上的噪声。而且不同模型间, 特征噪声的级别很难衡, 尤其是网络结构以及训练方法改变时, 度量变得困难, 去噪防御就会失效。本文提出一种基于局部邻域滤波的对抗攻击检测方法, 通过 RGB 空间切割得到相似块立方体, 并对立方体进行局部邻域滤波, 实现对抗攻击的检测防御。与已有方法相比, 本文方法不需要了解模型的参数, 也不需要知道模型的结构, 通过局部邻域来构造滤波参数, 对噪声进行度量, 根据噪声的级别进行滤波, 具有较高的通用性。

3 基于局部邻域滤波的检测方法

本文提出的方法主要包含两方面的工作, 分别是相似块分组滤波和一致性检验, 如图 2 所示。在相似块分组滤波阶段, 首先, 根据系数 d 对对抗样本(图 2a)下采样得到低分辨率样本(图 2b); 其次, 对低分辨率样本进行 RGB 空间切割和相似块分组(图 2c); 然后, 通过熵排序将相似块组成立方体(图 2d); 接着, 对立方体进行标准化处理, 得到标准立方体(图 2e); 最后, 通过标准立方体的三个邻域块得到邻域数据的标准差, 将其用于局部滤波降噪(图 2f)。在一致性检验阶段, 先将滤波后的相似块组逆标准化映射回 RGB 彩色空间, 组合得到滤波样本(图 2g); 再把原对抗样本和滤波样本分别输入模型, 得到分类结果, 进行一致性检验(图 2h)。

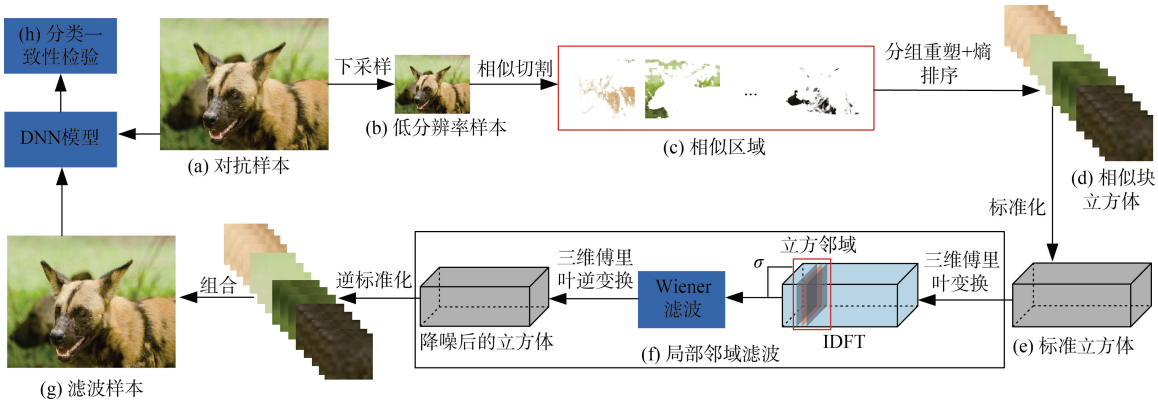

图 2 基于局部邻域滤波的检测框架图

Figure 2 Detection framework based on local neighborhood filtering

3.1 相似块分组

图 3 展示了相似块分组的基本流程。对于输入样本 x (图 3a), 先在 RGB 空间内对其进行不相邻相似块切割(图 3b), 得到相似像素点集合 M_i (图 3c); 再对 M_i 分组聚合, 获得相似图像块 m_{ij} (如图 3d)。其中, M_i 可表示为:

$$M_i = P_i\left(\mathrm{T}\left(x,K\right)\right),\ i=1,2,...,K \tag{1}$$

上式中, $\mathrm{T}(x,K)$ 在 RGB 空间内根据相似度将 x 的像素点分类聚合, 即把 x 的像素数据切割为 K 类, 每类中的像素可能不相邻, 但颜色差异小、相似度高。$P_i(\cdot)$ 是得到的第 i 类像素点集合的算子, 记作 M_i, 且 $H_i = |M_i|$ 表示 M_i 中的像素点数量, 若 x 的尺寸为 $U\times V\times 3$, 则 $U\times V = \sum_{i=1}^{K} H_i$ 。

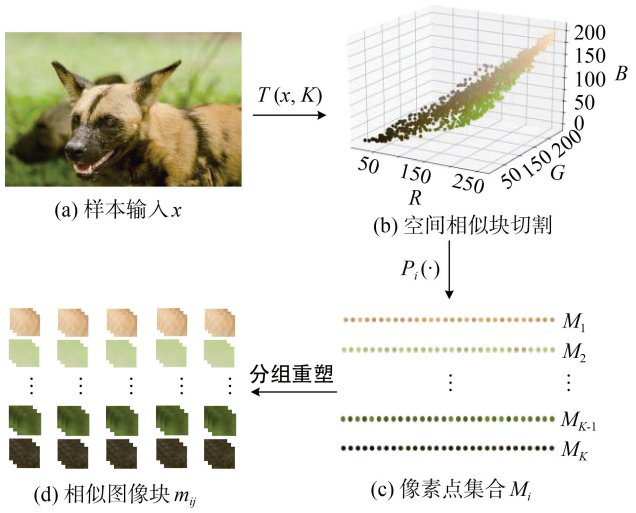

图 3 相似块分组框架

Figure 3 Similar block grouping framework

由于对抗样本中的噪声在不同灰阶上具有不同的强度, 本文方法在抠出来的像素集合 M_i 上构造适当尺寸的子块进行滤波。设所构造的子块图像(图 3d)尺寸为 $s\times s$, 则 $n_i = \left[H_i / s^2\right]$ 为 M_i 类上的子块数量, $[\bullet]$ 表示取整函数。设 $M_i = \left\{p_{i1}, p_{i2}, \cdots, p_{i,H_i}\right\} = P_i \cup T_i$, 其中 p_{ij} 是 M_i 中第 j 个像素的 RGB 灰阶的和, 且满足 $p_{ij} \leqslant p_{ik}$, $j \leqslant k$; P_i 和 T_i 分别称为 M_i 的主部和余部, 其中 P_i 由 $n_i s^2$ 个像素组成, 即:

$$P_i = \left\{p_{i1}, \cdots, p_{i,s^2}, \cdots, p_{i,(n_i-1)s^2+1}, \cdots, p_{i,n_i s^2}\right\} \tag{2}$$

为简便设 $P_i = \bigcup_{j=1}^{n_i} m_{ij}$, m_{ij} 表示 P_i 中的第 j 个子块, 则:

$$m_{ij} = \bigcup_{k=1}^{s^2} p_{i,(j-1)s^2+k} \tag{3}$$

本文将 m_{ij} 中 s^2 个元素聚合成图像子块, 即: 对于任意的 $p_{i,(j-1)s^2+k} \in m_{ij}$, 它在 m_{ij} 中的坐标 (r,c), $r,c = 1,2,\cdots,s$ 由(4)式确定。

$$\begin{cases} r = 1 + \left[\dfrac{k-1}{s}\right] \\ c = 1 + \mathrm{mod}(k-1, s) \end{cases}, \tag{4}$$

其中, $\mathrm{mod}(k-1,s)$ 表示 $k-1$ 模 s 的余数。

设所构造的相似块立方体为 u, 则相似块的构造如算法 1 所示。

算法 1: 相似块分组算法

输入: 图像 x, 下采样系数 d, 类参数 K, 聚合尺寸 s

输出: 相似块立方体 u

1. 根据 d 对 x 下采样
2. 通过 $T(x,K)$ 将 x 的 RGB 数据切割为 K 类
3. FOR $i = 1:K$ DO
4. 通过公式(1)的 $P_i(\cdot)$ 获取 M_i, $H_i = |M_i|$
5. 按公式(2)提取 M_i 的主部 P_i
6. 遍历 P_i 中元素, 按公式(4)将 P_i 分成 n_i 个子块 m_{ij}
7. END FOR
8. 将 m_{ij} 按字典排列组成 u
9. RETURN 相似块立方体 u

在构造相似块立方体 u 时, M_i 的余部 T_i 忽略不处理, 滤波操作仅在主部 P_i 上进行, 余部数据信息不变。

3.2 局部邻域滤波

在 Wiener 滤波处理前, 先对上一小节得到的相似块立方体 u (图 2d)的 RGB 值标准化(映射到 $[0,1]\times[0,1]\times[0,1] \subset R^3$ 空间), 得到标准立方体 Su (图 2e, 由标准化后的 m_{ij} 组合得到)。然后在标准立方体上进行局部邻域滤波, 这里的局部邻域滤波与寻常滤波的概念不同, 它主要是通过邻域立方体的 3 个块得到邻域数据的标准差(灰阶方差估计), 用于 Wiener 滤波降噪。当遇到不同对抗样本时, 立方体邻域内的统计特征会相应改变, 局部邻域滤波所使用的 3 维标准差也随之变化, 此时 Wiener 滤波能够灵

活地进行降噪, 从而实现对各类攻击的有效检测。

标准立方体的局部邻域滤波(图 2f, 参照算法 2)共分三步完成: a) 对标准立方体进行傅里叶变换; b) 在傅里叶空间逐层进行 Wiener 滤波; c) 进行傅里叶逆变换。

a) 三维傅里叶变换: 由于标准立方体中的图像块 m_{ij} 是三维数据, 而傅里叶变换通常在二维平面上进行, 因此本文使用空间分解, 先沿着 RGB 三通道将 m_{ij} 分解为 3 个二维数据。再对每个通道 c 上的二维数据进行傅里叶变换, 得到变换后的数据 dec_{ijc}, 其中 $c\in\{R,G,B\}$。最后, 按序将多个 dec_{ijc} 组合得到傅里叶变换体 $IDFT$。

b) 逐层 Wiener 滤波: 图像块 m_{ij} 上的噪声是独立同分布的加性噪声, 为有效减轻这些噪声的影响, 本文在 $IDFT$ 中逐层进行 Wiener 滤波。其主要思想是依次在傅里叶变换体的某一帧及其上下邻域帧(共 3 帧)上构造灰阶方差估计 σ_{ijc}/e (e 是一个整数值), 并将该方差估计用于 dec_{ijc} 上的 Wiener 滤波, 破坏扰动噪声, 得到新的数据 dec_{ijc}^{w}。

c) 三维傅里叶逆变换: 为将滤波降噪后的 dec_{ijc}^{w} 重组为空域立方体, 需对其进行三维傅里叶逆变换。三维傅里叶逆变换是先根据 RGB 三通道对 dec_{ijc}^{w} 进行二维傅里叶逆变换; 再沿着第三维合并 RGB 三个通道上的数据, 得到滤波降噪后“干净”的新图像块 $m_{ij}^{'}$; 最后按序排列 $m_{ij}^{'}$, 得到降噪后的立方体 Fu。

根据上述步骤, 标准立方体的局部邻域滤波如算法 2 所示。

算法 2: 局部邻域滤波算法

输入: 类参数 K, 聚合尺寸 s, 标准立方体 Su 的图像块 m_{ij}, 整数 e

输出: 降噪后的立方体 Fu

1. 将 m_{ij} 分解为二维数据 $m_{ijc}(c\in\{R,G,B\})$
2. 遍历 Su, 对 m_{ijc} 进行傅里叶变换得到 $IDFT=\{dec_{ijc}\}$
3. FOR $i=1:K$ DO
4. FOR $j=1:n_i$ DO
5. 用 $IDFT_{ij}$ 帧及其上下邻域帧构造 σ_{ijc}/e
6. 将 σ_{ijc}/e 用于 Winer 滤波, 得到 dec_{ijc}^{w}
7. 对 dec_{ijc}^{w} 进行傅里叶逆变换得到 $m_{ijc}^{'}$, 根据 RGB 三通道合并 $m_{ijc}^{'}$ 得到 $m_{ij}^{'}$
8. END FOR
9. END FOR
10. $Fu=\{m_{ij}^{'}\}$。
11. RETURN 降噪后的立方体 Fu

3.3 一致性检验

以上步骤完成后, 将降噪后的立方体(即相似图像块组)逆标准化映射回 RGB 彩色空间, 并组合得到滤波样本 x' (如图 4 所示)。然后将原样本 x 和滤波样本 x' 输入分类器, 得到分类结果分别为 l_x 和 $l_{x'}$, 进行一致性检验(图 2h), 如果分类结果相同 $l_x=l_{x'}$, 则判定为正常样本 $y_x=1$, 反之, 则判定为对抗样本 $y_x=0$。即:

$$y_x=\begin{cases}1, & l_x=l_{x'}\\ 0, & l_x\neq l_{x'}\end{cases}。\tag{5}$$

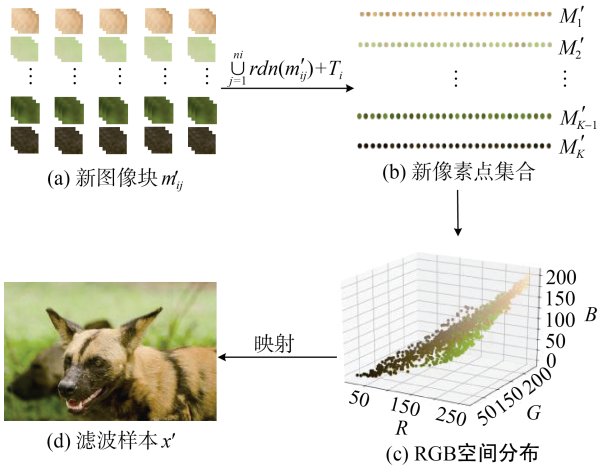

图 4 图像组合框架

Figure 4 Image composition framework

整个算法如算法 3 所示。

算法 3: 对抗攻击的检测算法

输入: 下采样系数 d, 类参数 K, 聚合尺寸 s, 图像 x

输出: 检测结果 y_x

1. 根据算法 1 对 x 分组, 得到相似块立方体 u
2. 标准化 u, 得到标准立方体 Su
3. 由算法 2 获得降噪后的立方体 Fu
4. FOR $i=1:K$ DO
5. $M_i^{'}=\bigcup_{j=1}^{n_i} rdn(m_{ij}^{'})+T_i$

6. 将 $M_i^{'}$ 映射回原位置

7. END FOR

8. 得到 x', 并将 x' 和 x 输入分类器, 根据公式(5)进行一致性检验

9. RETURN 检测结果 y_x

组合滤波样本 x' 时, 需将滤波后的图像块 $m_{ij}^{'}$ 逆重塑回像素点集合, 与未处理的余部数据 T_i 合并成 $M_i^{'}$。由于 T_i 保存的数据是 M_i 中的最后一部分像素, 只需按照顺序合并 rdn($m_{ij}^{'}$) 和 T_i, 就可得到像素点集合 $M_i^{'}$, 其中, rdn(·) 是将立方体中的 $m_{ij}^{'}$ 逆重塑回像素点集合。最后根据 $M_i^{'}$ 的位置标记, 将 $M_i^{'}$ 中的像素点按照顺序映射回其在图像中的原始位置(如图 4 所示), 组合得到滤波样本 x'。

4 实验

本文使用 mini-ImageNet 分类数据集去评估基于局部邻域滤波的检测方法。该数据集节选自 ImageNet 数据集, 包含 100 类共 60k 张彩色图片。在实验中, 本文只考虑非定向的白盒攻击, 且每种攻击择取的图片数量不少于 2k 张。本文使用 ResNet-50、Inception-v3、ResNet-101 以及 VGG-16 作为目标分类器, 并使用预先训练的模型进行攻击和防御实验。

根据对抗攻击的定义, 如果良性样本已经被错误分类, 默认情况下被视为攻击成功。此时, 攻击者不用进行攻击操作, 只需返回未经修改的良性样本。因此, 本文的实验仅限于正确分类的良性样本。实验中的攻击模型来自于 torchattacks 库, 该库包含 FGSM、DeepFool、BIM、C&W、PGD 等经典攻击方法。

4.1 参数的消融实验

本文的检测模型有三个超参数: 分类数量 K, 下采样系数 d 以及 Wiener 滤波系数 σ_{ijc}/e。它们分别控制着相似块分组的数量、下采样后图片的尺寸大小以及滤波降噪的程度。其数值过大或过小会损害良性样本的特征信息, 影响模型对良性样本的识别精度。因此, 需要选择合适的超参数值, 本文综合在对抗样本和良性样本上的检测精度, 对上述 3 个参数进行消融实验。

本文从数据集中择选 2k 张图片来分析不同的超参数值对检测结果的影响, 使用 ResNet-50 目标分类器以及 FGSM 攻击(参数 $\varepsilon = 4/255$)进行各个参数的评估。

为择取某一参数的最优值, 本文将其他两个超参数设为指定值, 仅在剩余的一个超参数上寻找最大性能的设置, 实验结果如表 1 所示。表 1 的第一列是超参数的多组配置, 其余四列是良性样本的分类精度。

表 1 超参数的消融实验

Table 1 Ablation experiments on hyperparameters

(K, d, e)	ResNet-50	Inception-v3	ResNet-101	VGG-16	平均精度
(8, 0.5, 4800)	0.903	0.906	0.912	0.852	0.893
(8, 0.5, 4900)	0.913	0.910	0.915	0.870	0.902
(8, 0.5, 5000)	0.917	0.920	0.923	0.880	0.910
(8, 0.6, 5000)	0.930	0.927	0.931	0.895	0.921
(8, 0.7, 5000)	0.940	0.939	0.938	0.910	0.932
(9, 0.7, 5000)	0.954	0.942	0.939	0.916	0.938
(10, 0.7, 5000)	0.955	0.946	0.941	0.919	0.940

由表 1 可知, 随着 K 值、e 值和 d 值的增加, 良性样本的分类精度逐渐提高, 并在 $K=10$、$d=0.7$、$e=5000$ 时趋于平稳, 此时平均分类精度达到 0.940。除此之外, 对于各组参数配置, VGG-16 的分类精度都要低于其他三个模型, 这表明了 VGG-16 对去噪操作的接受度低于 ResNet-50、Inception-v3 和 ResNet-101。

通过查看表 1 的分类精度和多次实验中的防御效果, 本文发现在 $K=10$、$d=0.7$、$e=5000$ 时, 基于局部邻域滤波的检测方法既可以取得较好的检测效果, 也能够保留良性样本的数据信息。且从图 5 中可观察到对抗样本在通过本文方法处理后, 其包含的扰动已被滤除, 成功防御了对抗攻击。本文采用检测率(DNR)来评价本文方法, 如公式(6)所示。

$$\mathrm{DNR} = n_{adv}^{detect} \Big/ n_{adv} \tag{6}$$

其中 n_{adv} 表示攻击成功的对抗样本数; n_{adv}^{detect} 表示被成功检测的攻击样本数量。

4.2 检测方法的性能

为验证本文检测方法的通用性, 本文采用多种攻击技术去生成对抗样本, 并用 4.1 小节择取的超参数来检测图像是否被攻击。实验中, 我们从 mini-ImageNet 数据集中抽取 2k 张图像进行防御测试; 这些图像包含 40 个类别, 且每类至少有 40 个样本。对于无攻击时分类错误的图像, 本文没有进行防御测试。

表 2 显示了本文方法在 FGSM、BIM、DFool、C&W、PGD、MI-FGSM、FFGSM、TPGD、RFGSM 攻击下的 DNR。实验模型包括 Inception-v3、VGG-16、ResNet-50 以及 ResNet-101, 表 2 中数据由公式(6)进行 DNR 统计得到。

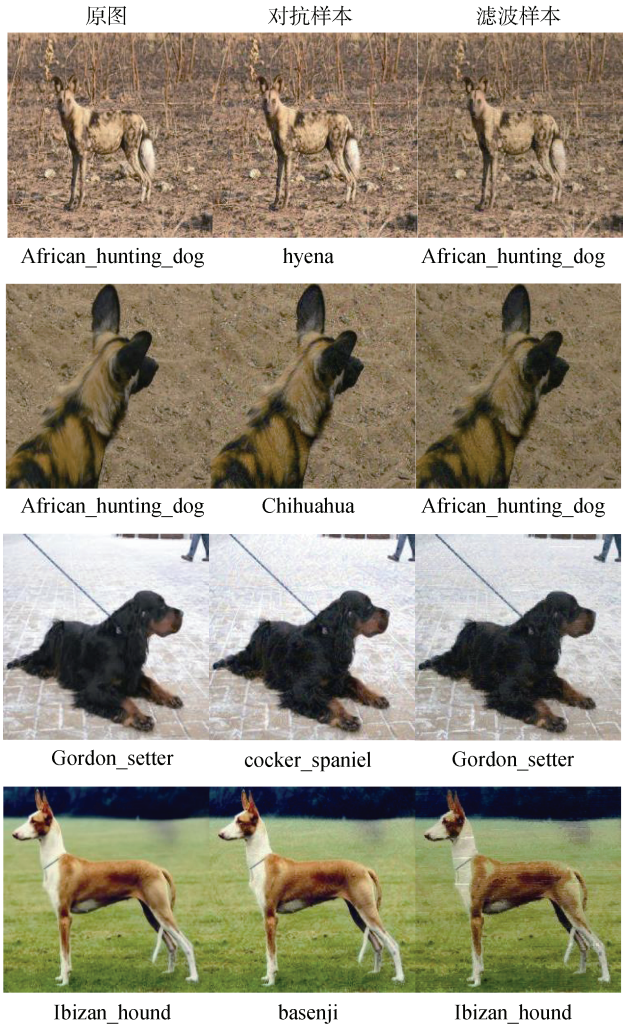

图 5 检测攻击的样图

Figure 5 Sample drawing of detection attack

表 2 本文方法在不同攻击下的 DNR

Table 2 DNR of our method under different attacks

Attack	ResNet-50	Inception-v3	ResNet-101	VGG-16
FGSM	0.860	0.770	0.866	0.848
BIM	0.821	0.674	0.840	0.679
DFool	0.884	0.869	0.893	0.890
C&W	0.913	0.925	0.938	0.888
PGD	0.924	0.855	0.928	0.899
MI-FGSM	0.877	0.765	0.889	0.832
FFGSM	0.859	0.783	0.871	0.852
TPGD	0.922	0.862	0.919	0.867
RFGSM	0.876	0.785	0.879	0.864

表 2 表明本文的检测方法在不同模型下对各个攻击都有良好的检测效果, 可应用于多种分类模型。从表中的检测数据可得出, 相同实验环境设置中, 在 ResNet-101 分类模型下的检测率均高于 Inception-v3 和 VGG-16; 此外, 在 ResNet-50 分类模型下, 针对 PGD 攻击的检测率高于其他类型的攻击。且从整体来看, 本文方法在 DFool、C&W、PGD 以及 TPGD 攻击下都取得了较好的检测结果, 是一种不依赖攻击方法的通用防御策略, 具有较强的通用性。

4.3 实验结果对比

在实施攻击时, 不同的作者会设置不同的攻击参数和扰动幅度, 导致很难平衡攻击的强度[30-31]。于是, 为保证公平性, 本文统一使用 DNR 来衡量各个防御方法, 其值为 1 表示在防御处理后, 攻击导致分类错误的对抗样本都被检测出来。

表 3 比较了本文方法与其他防御模型在 ResNet-50 下的检测率。其中, 第一列是防御模型, 其余两列是防御 FGSM 和 C&W 攻击的 DNR。FGSM 的攻击参数 $\varepsilon = 4/255$, C&W 采用 $steps = 1000$ 的 L_2 无目标攻击。

表 3 ResNet-50 模型下的 DNR

Table 3 DNR under ResNet-50 model

防御方法	ResNet-50	
	FGSM	C&W
(a) Saliency Map(θ =0.5) [32]	0.833	0.772
(a) + Color Reversing(θ =0.1) [32]	0.759	0.478
(a) + Zero Mean(θ =0.1) [32]	0.892	0.728
本文方法	0.860	0.913

表 3 表明, 本文方法与 Saliency Map[32]相比, 针对 C&W 攻击, 本文方法取得了较优的检测结果。此外, 尽管本文方法针对 FGSM 攻击的检测效果不如 Zero Mean 方法, 但比 Saliency Map 和 Color Reversing 的防御效果都要好。

表 4 显示了本文方法与 RBF-SVM、FS、NR 在 VGG-16 模型下的对比结果。其中, 第一列是防御模型, 其余三列是针对 FGSM、C&W 和 DFool 攻击的 DNR。且 FGSM 的攻击参数 $\varepsilon = 4/255$, DFool 采用 $steps = 3$ 的 L_2 无目标攻击, C&W 攻击的配置和表 3 相同。

表 4 VGG-16 模型下的 DNR

Table 4 DNR under VGG-16 model

防御方法	VGG-16		
	FGSM	C&W	DFool
RBF-SVM[13]	0.826	0.519	0.601
FS[17]	0.286	0.893	0.744
NR[27]	0.775	**0.923**	**0.921**
本文方法	**0.848**	0.888	0.890

表 4 表明, 针对 FGSM、C&W 与 DFool 攻击, 本文方法的检测率分别为 0.848、0.888 和 0.890, 都优于 RBF-SVM 方法。且相比于 FS 防御模型, 本文方

法在 FGSM 和 DFool 攻击下也都取得了较优的检测结果。此外, 针对 FGSM 攻击的检测结果, 本文方法也是优于 NR 防御模型。

表 5 显示了本文方法与多种防御模型在 Inception-v3 下的对比结果。其中, 第一列是防御模型, 其余三列是针对 FGSM、MIM、PGD 和 BIM 攻击的 DNR, 且这四种攻击的参数 ε 都为 4/255 。

表 5 Inception-v3 模型下的 DNR

Table 5 DNR under Inception-v3 model

Defense	FGSM	MIM	PGD	BIM
G-RGB[16]	0.134	0.140	0.116	0.108
GTS[33]	0.200	0.210	0.232	0.224
KD+BU[34]	0.574	0.553	0.574	0.460
RGB[33]	0.534	0.548	0.935	0.202
Noise[33]	0.664	0.670	0.972	0.460
TSD[33]	0.707	0.699	**0.984**	0.507
本文方法	**0.770**	**0.765**	0.855	**0.674**

表 5 展示了在 Inception-v3 模型下, 各类防御机制在对抗样本上的检测率。从表中可以看出, 本文方法在 FGSM、MIM 以及 BIM 攻击下取得了较高的检测率, 积极防御了对抗样本的攻击。

Prakash 等[30]发现对抗样本给出的错误预测经常出现在良性样本的前五预测中, 近 40%的对抗样本是原良性样本分类的第二最可能预测类, 原始预测结果通常依然保留在对抗样本的前五预测中。本文方法通过破坏对抗样本中的扰动, 使前五预测的排名改变, 从而检测出对抗样本。从表 3、表 4 和表 5 中也可发现本文方法在 FGSM、C&W 等攻击下都取得了较优的检测结果, 有效拒绝了对抗样本的输入, 保障了分类模型的安全。

5 结论

为了提高分类网络模型防御对抗样本的能力, 本文提出了基于局部邻域滤波的对抗攻击检测方法, 对各类攻击方法都具有通用性。在检测过程中, 我们将相似的图像块组成立方体, 并基于立方体中的邻域进行局部滤波。此时, 当不同类型的对抗攻击训练出对抗扰动时, 立方体邻域内的统计特征会相应改变, 从而 Wiener 滤波能够有效地检测出对抗样本。实验结果表明, 本文的检测方法具有较高的检测精度。本文提出的方法不需要修改网络模型, 具有应用简单的优势。相较于 UIPD 等基于迭代优化的防御手段, 本文方法在模型间的泛化能力较强, 可以跨模型防御对抗样本的攻击, 有效确保了分类结果的可信性。

参考文献

[1] Yang J, Dang J S. Semantic Segmentation of 3D Point Cloud Based on Contextual Attention CNN[J]. *Journal on Communications*, 2020, 41(7): 195-203.
(杨军, 党吉圣. 基于上下文注意力 CNN 的三维点云语义分割[J]. *通信学报*, 2020, 41(7): 195-203.)

[2] Tian C, Wang L, Zhou E F, et al. Integration and Experimental Study of Automatic Driving System for Bus[C]. *2021 7th International Symposium on Mechatronics and Industrial Informatics*, 2021: 96-103.

[3] Li D J, Li D G, Yang L. Application of Convolutional Neural Network in Dynamic Gesture Tracking[J]. *Journal of Frontiers of Computer Science and Technology*, 2020, 14(5): 841-847.
(李东洁, 李东阁, 杨柳. 卷积神经网络在动态手势跟踪中的应用[J]. *计算机科学与探索*, 2020, 14(5): 841-847.)

[4] Huang Y L, Luo X X, Liu D R. Local Finite Weight Sharing of MFSC Coefficients Based CNN Speech Recognition[J]. *Control Engineering of China*, 2017, 24(7): 1507-1513.
(黄玉蕾, 罗晓霞, 刘笃仁. MFSC 系数特征局部有限权重共享 CNN 语音识别[J]. *控制工程*, 2017, 24(7): 1507-1513.)

[5] Goodfellow I J, Shlens J, Szegedy C. Explaining and Harnessing Adversarial Examples[EB/OL]. 2014: arXiv: 1412.6572. https://arxiv.org/abs/1412.6572.

[6] Kurakin A, Goodfellow I J, Bengio S. Adversarial Examples in the Physical World[M]. Artificial Intelligence Safety and Security. First edition. | Boca Raton, FL: CRC Press/Taylor & Francis Group, 2018.: Chapman and Hall/CRC, 2018: 99-112.

[7] Carlini N, Wagner D. Towards Evaluating the Robustness of Neural Networks[C]. *2017 IEEE Symposium on Security and Privacy*, 2017: 39-57.

[8] Papernot N, McDaniel P, Goodfellow I. Transferability in Machine Learning: From Phenomena to Black-Box Attacks Using Adversarial Samples[EB/OL]. 2016: arXiv: 1605.07277. https://arxiv.org/abs/1605.07277.

[9] Xiao C W, Li B, Zhu J Y, et al. Generating Adversarial Examples with Adversarial Networks[C]. *The 27th International Joint Conference on Artificial Intelligence*, 2018: 3905-3911.

[10] Moosavi-Dezfooli S M, Fawzi A, Frossard P. DeepFool: A Simple and Accurate Method to Fool Deep Neural Networks[C]. *2016 IEEE Conference on Computer Vision and Pattern Recognition*, 2016: 2574-2582.

[11] Irfan M M, Ali S, Yaqoob I, et al. Towards Deep Learning: A Review on Adversarial Attacks[C]. *2021 International Conference on Artificial Intelligence*, 2021: 91-96.

[12] Dong Y P, Liao F Z, Pang T Y, et al. Boosting Adversarial Attacks with Momentum[C]. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018: 9185-9193.

[13] Lu J J, Issaranon T, Forsyth D. SafetyNet: Detecting and Rejecting Adversarial Examples Robustly[C]. *2017 IEEE International Conference on Computer Vision*, 2017: 446-454.

[14] 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]. *The 10th ACM Workshop on Artifi-*

cial Intelligence and Security, 2017: 15-26.

[15] Yao Z W, Gholami A, Xu P, et al. Trust Region Based Adversarial Attack on Neural Networks[C]. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019: 11350-11359.

[16] Metzen J H, Genewein T, Fischer V, et al. On Detecting Adversarial Perturbations[C]. *2017 International Conference on Learning Representations*, 2017.

[17] Yan F, Zhang M L, Zhang L Q. Adversarial Examples Detection Method Based on Boundary Values Invariants[J]. *Chinese Journal of Network and Information Security*, 2020, 6(1): 38-45.
(严飞, 张铭伦, 张立强. 基于边界值不变量的对抗样本检测方法[J]. *网络与信息安全学报*, 2020, 6(1): 38-45.)

[18] 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*, 2018: 15-26.

[19] Pang T, Du C, Zhu J. Robust Deep Learning via Reverse Cross-Entropy Training and Thresholding Test[J]. 2017: ArXiv Preprint ArXiv:1706.00633, 3.

[20] Meng D Y, Chen H. MagNet: A Two-Pronged Defense Against Adversarial Examples[C]. *The 2017 ACM SIGSAC Conference on Computer and Communications Security*, 2017: 135-147.

[21] Ding X T, Cheng Y Q, Luo Y L, et al. Consensus Adversarial Defense Method Based on Augmented Examples[J]. *IEEE Transactions on Industrial Informatics*, 2023, 19(1): 984-994.

[22] Hinton G, Vinyals O, Dean J. Distilling the Knowledge in a Neural Network[C]. NIPS 2014 Deep Learning Workshop, 2014.

[23] Chen J Y, Wu C A, Zheng H B, et al. Universal Inverse Perturbation Defense Against Adversarial Attacks[J]. *Acta Automati-Ca Sinica*, 2021: 1-16.
(陈晋音, 吴长安, 郑海斌, 等. 基于通用逆扰动的对抗攻击防御方法[J]. *自动化学报*, 2021: 1-16.)

[24] Xie C H, Wu Y X, van der Maaten L, et al. Feature Denoising for Improving Adversarial Robustness[C]. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019: 501-509.

[25] Song Y, Kim T, Nowozin S, et al. PixelDefend: Leveraging Generative Models to Understand and Defend Against Adversarial Examples[C]. *2018 International Conference on Learning Representations*, 2018.

[26] Samangouei P, Kabkab M, Chellappa R. Defense-GAN: Protecting Classifiers Against Adversarial Attacks Using Generative Models[C]. *2018 International Conference on Learning Representations*, 2018.

[27] 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.

[28] Wu F, Yang W X, Xiao L M, et al. Adaptive Wiener Filter and Natural Noise to Eliminate Adversarial Perturbation[J]. *Electronics*, 2020, 9(10): 1634.

[29] Gupta P, Rahtu E. CIIDefence: Defeating Adversarial Attacks by Fusing Class-Specific Image Inpainting and Image Denoising[C]. *2019 IEEE/CVF International Conference on Computer Vision*, 2019: 6708-6717.

[30] Prakash A, Moran N, Garber S, et al. Deflecting Adversarial Attacks with Pixel Deflection[C]. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018: 8571-8580.

[31] Xie C H, Wang J Y, Zhang Z S, et al. Mitigating Adversarial Effects through Randomization[C].2018 *International Conference on Learning Representations*, 2018.

[32] Ye D P, Chen C X, Liu C R, et al. Detection Defense Against Adversarial Attacks with Saliency Map[J]. *International Journal of Intelligent Systems*, 2022, 37(12): 10193-10210.

[33] Gao S, Yu S, Wu L W, et al. Detecting Adversarial Examples by Additional Evidence from Noise Domain[J]. *IET Image Processing*, 2022, 16(2): 378-392.

[34] Feinman R, Curtin R R, Shintre S, et al. Detecting Adversarial Samples from Artifacts[EB/OL]. 2017: arXiv: 1703.00410. https://arxiv.org/abs/1703.00410.

刘朝 于 2019 年在常州大学计算机科学与技术专业获得学士学位。现在安徽师范大学计算机科学与技术专业攻读硕士学位。研究领域为: 图像处理、深度学习、人工智能安全。

朱莉芳 于 2021 年在安徽师范大学软件工程专业获得学士学位。现在安徽师范大学计算机科学与技术专业攻读硕士学位。研究领域为: 图像处理、深度学习、人工智能安全。

接标 于 2015 在南京航空航天大学获得学博士学位。现任安徽师范大学计算机与信息学院教授。研究领域为: 机器学习, 模式识别, 数据挖掘等。

丁新涛 于 2015 在安徽师范大学获得学博士学位。现任安徽师范大学计算机与信息学院副教授。研究领域为: 机器学习、计算机视觉等。