

## 联合图像生成和图像重构的对抗样本检测方法<sup>①</sup>

李宝平<sup>②</sup> 夏瑜昊<sup>③</sup>

(河南理工大学物理与电子信息学院 焦作 454000)

**摘要** 对抗样本攻击是识别网络面临的主要安全威胁之一。针对对抗样本检测过程中由分类边界模糊导致识别准确率低及需大量对抗样本参与训练导致模型收敛速率慢等问题,本文提出一种联合图像重构技术和图像生成技术实现的对抗样本检测方法。首先,设计由卷积层和 Swin-Transformer 联合实现的图像重构网络,还原图像的语义信息并消除对抗性扰动;然后,利用条件生成式对抗网络依据标签信息生成对应类别图像;最后,将重构图像和生成图像送至卷积识别网络,依据分类结果一致性判断是否为对抗样本。该方法将对对抗样本检测问题转化为图像分类问题,无需对抗样本参与训练,无需先验地了解攻击者的攻击类型和被攻击模型的结构和参数即可直接检测对抗样本。在 VGG-16、ResNet-18、GoogLeNet 分类网络和 MNIST、GTSRB 数据集上的实验结果表明,该检测方法相较于其他经典检测方法,平均识别准确率提升了 4.75%~22.86%,F1 分数提升了 3.40%~13.64%,证明了其优越性。

**关键词** 对抗样本; 图像分类; Swin-Transformer; 图像重构; 卷积神经网络

随着计算机算力的持续增强以及来自各领域海量数据的深度整合,深度神经网络(deep neural network, DNN)<sup>[1]</sup>在许多具有挑战性的任务中取得了瞩目的成就,例如人脸识别、图像分类、无人驾驶等任务。但研究表明,深度神经网络自身存在缺陷,深度学习的不可解释性导致其输出缺乏可信度并且容易遭受对抗攻击<sup>[2]</sup>,输出错误的预测或分类结果。因此,在将深度学习模型应用于诸如汽车自动驾驶<sup>[3]</sup>、医学诊断、安全监控等敏感场景时,必须着重考虑模型的安全性和可靠性<sup>[4]</sup>。文献[5]提出了对抗样本的概念,指出深度学习模型在面对对抗样本时表现得非常脆弱。例如,在文献[3]中,攻击者在交通指示牌上恶意添加一些难以察觉的扰动性噪声,可能导致汽车在自动驾驶时将“停止”标志识别为“限速”,进而引发严重的安全事故。这种脆弱性印证了深度学习模型在处理对抗样本时的局限性,

也促使研究人员寻找更具鲁棒性的模型和防御方法,以应对对抗样本攻击带来的挑战。

目前,针对对抗样本的防御方法分为主动型方法和反应式方法。其中,主动型方法是针对 DNN 模型进行强化,使得模型在对抗攻击中仍能保证正确的分类结果,但该方法容易被针对性训练,导致防御性能下降。反应式方法则是依据输入的正常样本与对抗样本的特征差异进行检测识别,其检测过程与识别网络无关,具有较好的普适性和鲁棒性,是现阶段比较主流的检测方法,也被称为对抗样本检测算法。

目前对抗样本检测方法可分为 3 类:(1)通过对抗样本和正常样本在统计学中的特征差异性检测对抗样本;(2)在 DNN 模型内部或外部引入检测器模块检测对抗样本;(3)通过预测结果的不一致性检测对抗样本。本文所提方法正是利用图像语义信

① 国家自然科学基金(62101176)资助项目。

② 男,1981 年生,博士,副教授;研究方向:智能信号处理;E-mail: libaoping@hpu.edu.cn。

③ 通信作者,E-mail: 1056350048@qq.com。

(收稿日期:2024-07-24)

息和标签信息的不一致性检测对抗样本。

文献[6]基于条件像素卷积神经网络(convolutional neural network, CNN)模型的条件生成能力和对图像序列的逐像素预测机制重构输入图像,分析输入样本的每个像素值,并考虑其前面所有的像素值及输入的条件信息(如类别标签)。通过比较输入样本与模型预测像素值之间的差异性实现对抗样本检测,该算法所需的收敛耗时较长。文献[7]提出了使用75%随机掩码和Swin-Transformer的图像重构方法,虽然随机掩码可以较好抑制对抗噪声,但同时也会损失大部分细节,使图像识别准确率降低。

文献[8]使用压缩样本特征输入空间的方法实现对抗样本检测。通过对原始样本进行特征压缩,将不同特征向量对应的样本合并成一个样本,比较模型对原始输入的预测与对实施特征压缩后输入的预测结果,实现对抗样本检测。该方法对于某些攻击算法生成的扰动较大的对抗样本检测效果不够理想。

文献[9]使用基于对抗特征统计的方法检测对抗样本,对抗性检测网络(adversary detection network, ADN)没有直接针对原始样本进行检测,而是固定原始分类网络的权重,获取其中间层的输入并采用交叉熵作为损失函数训练检测网络。该方法需使用大量对抗样本和正常样本同时参与训练,太过依赖已知的对抗样本,且训练耗时较长。

针对上述文献方法中存在的检测准确率低、收敛速率慢且过于依赖潜在攻击的先验知识等问题,本文提出了一种联合图像重构技术<sup>[10]</sup>和图像生成技术<sup>[11]</sup>实现的对抗样本检测方法(reconstitution-generation, Recon-Gen)。该方法主要基于以下思考:为了使对抗性扰动不易被人类视觉系统所察觉,对抗攻击引起的扰动通常被限制在一个小范围内,且这些细微扰动使得图像被神经网络错误地识别为其他类别,即对抗性扰动的引入导致了图像语义信息和标签信息的不一致性。将这种对抗性扰动视为一种噪声,并采用图像重构网络对包含对抗性噪声的图像进行降噪,若扰动的效果被适当降级,则图像会被分类为其他类别。而正常图像经过降噪仍保持原来的分类类别,以此区分对抗样本与正常样本。

本文贡献总结如下。

(1) 本文所提检测方法实现了将对抗样本检测问题转化为图像分类问题。该方法无需对抗样本参与训练,仅使用正常样本训练即可。且该方法具有良好的泛化性能,无须先验地了解攻击者的攻击类型和被攻击模型的具体结构和参数即可直接检测对抗样本。

(2) 在图像重构阶段,设计了一种基于卷积层和Swin-Transformer混合架构的图像重构方法,并在深层特征提取模块中采用滑动窗口式自注意力机制,该机制可以实现窗口之间的信息交互,有助于神经网络学习到图像的全局特征,易于检测出包含细微扰动的对抗样本。

(3) 在图像生成阶段使用条件生成式对抗网络(conditional Wasserstein generative adversarial network-gradient penalty, CWGAN-GP)依据标签类别直接生成完整图像,无需分析输入样本的每个像素值进而逐像素点地生成图像,可有效缩减网络模型规模和训练耗时。图像生成过程可看作是一个“升维”过程,用以消除低维标签信息和高维图像信息之间存在的匹配模糊问题,提升识别准确率。

## 1 相关工作

快速梯度攻击法(fast gradient sign method, FGSM)<sup>[12]</sup>沿着损失函数梯度上升的方向对图像添加扰动,生成对抗样本。FGSM的优势在于其简单性和高效性,只需计算1次梯度即可生成对抗样本。可用公式表示为

$$x' = x + \varepsilon \cdot \text{sign}[\nabla_x J(\theta, x, y)] \quad (1)$$

式中: $x$ 为原始样本; $x'$ 为添加扰动后的对抗样本; $\nabla_x$ 是损失函数相对于输入 $x$ 的梯度; $J(\theta, x, y)$ 即损失函数, $\theta$ 是模型的权重参数, $y$ 是 $x$ 的真实类别; $\text{sign}()$ 是符号函数,将输入转换为-1和+1两类符号输出; $\varepsilon$ 为扰动强度。

迭代攻击法(basic iterative method, BIM)<sup>[13]</sup>基于FGSM方法进行改进,BIM沿损失函数梯度上升的方向多次添加扰动,并在每次迭代之后都对所得结果的像素值进行裁剪,保证结果处于 $\varepsilon$ 邻域范围

内。该方法生成的对抗样本可描述为

$$x_{i+1} = clip_{\epsilon} \{ x_i + \alpha \cdot \text{sign}(\nabla_x L(x_i, y)) \} \quad (2)$$

式中： $i$  为迭代次数； $x_0 = x$  为正常输入； $\alpha$  表示迭代的步长参数； $clip(\cdot)$  为裁剪函数。

C&W 攻击 (Carlini & Wagner's attack)<sup>[14]</sup> 使用自适应矩估计优化器生成对抗样本, 引入变量  $w_n \in [-\infty, +\infty]$ , 利用映射函数将对抗样本的像素点取值映射到  $[-1, +1]$  区间, 以满足箱型约束条件, 则扰动  $r_n$  可表示为

$$r_n = \frac{1}{2}(\tanh(w_n) + 1) - x \quad (3)$$

定义 C&W 的损失函数为

$$\min_{w_n} r_n + c \cdot f\left(\frac{1}{2}(\tanh(w_n) + 1)\right) \quad (4)$$

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

式中： $c$  为超参数, 控制添加的扰动值与错误分类置信度之间的平衡； $Z(x_{adv})$  表示对抗样本  $x_{adv}$  的输出向量； $k$  表示置信度,  $k$  越大说明模型分类错误的概率越大。

DeepFool<sup>[15]</sup> 是一种基于超平面分类的攻击方

法, 其关键步骤是找到在哪个方向上对输入添加扰动, 可以有效地改变模型的预测。为此, 它通过计算一个“最小扰动”的量来实现。这个量表示使模型的预测从原始类别变为另一个类别的最小扰动大小, 但该算法复杂度较高。

## 2 对抗样本检测方法实现

### 2.1 总体设计

Recon-Gen 系统的总体架构如图 1 所示。首先, 通过图像重构, 还原图像中包含的深层语义特征, 去除图像中可能存在的对抗性扰动; 然后, 利用条件生成式对抗网络依据文本标签生成对应类别图像, 将低维标签信息转化为高维图像信息, 以解决低维标签信息和高维图像信息之间存在的匹配模糊问题; 最后, 将重构图像与生成图像同时输入至分类网络中进行标签“一致性”检测, 即对比图像的语义信息与标签信息是否一致, 实现对抗样本检测。该检测过程将对抗样本检测问题转化为图像分类问题, 在无需了解被攻击模型参数的情况下即可实现对抗样本检测。

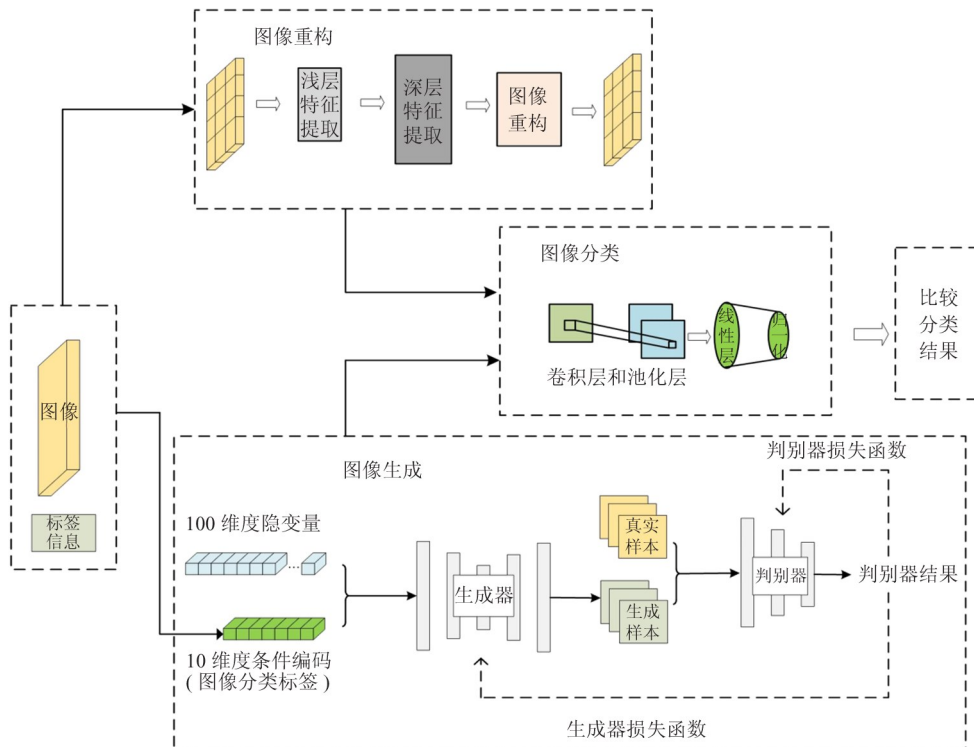


图 1 Recon-Gen 检测方法架构

## 2.2 图像重构

本文将对抗性扰动视为一种噪声,并采用基于滑动窗口自注意力机制的重构技术对图像进行降噪,若对抗性扰动的效果被适当降级,则降噪后的对抗样本将被分类为一个新类别。而对于正常样本,由于对抗性扰动是后续添加到图像表面的,相较于正常图像,扰动对降噪过程的耐受度更低,相同的降噪过程不会改变图像原有的语义信息,仍保持原先的分类类别。

近年来, Swin-Transformer<sup>[16]</sup> 在图像重构领域取得了瞩目的成就。一方面, Swin-Transformer 在每个分块内部引入了多头注意力机制,可以同时关注不同尺度的特征;这种多尺度特征融合的特性有助于模型更好地理解图像的语义信息,并在图像重构过程中实现更加细致和准确的重建。另一方面,它可以通过滑动窗口的方法建立长距离依赖模型,经过如图 2 所示的变换后,可以实现不同窗口之间的信息交互,并通过不断合并图像中的像素块以降低噪声的复杂度。由此考虑采用基于 Swin-Transformer 架构的 SwinIR (Swin-Transformer for image restoration) 强线性神经网络实现重构输入图像,去除对抗性扰动。

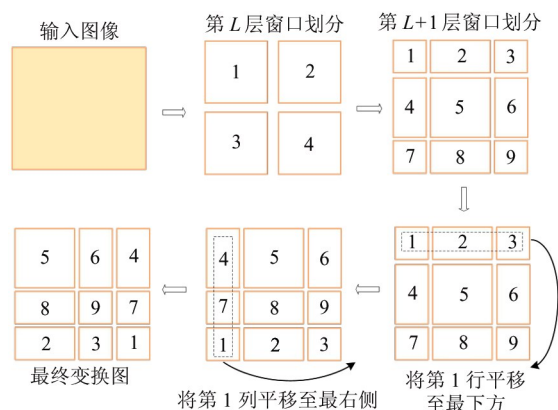


图 2 滑动窗口变换示意图

改进后的 SwinIR 网络架构如图 3 所示,在其深层特征提取模块中采用了滑动窗口式自注意力机制计算。在处理每个窗口时,自注意力机制被应用于窗口内的所有像素块,以捕捉局部区域间的特征关系。与此同时,为了降低噪声复杂度并提高计算效率,将像素块逐渐合并,使得每个窗口的特征表达更加紧凑。这种方法充分利用了自注意力机制的能力来捕获图像中的重要信息,并通过合并像素块来简化处理,从而使得重构网络更好地学习到图像中深层次的语义特征,有效消除图像中的对抗性噪声。

首先,给定一个输入  $I_{LQ} \in R^{H \times W \times C_{in}}$  ( $H$ 、 $W$  和  $C_{in}$

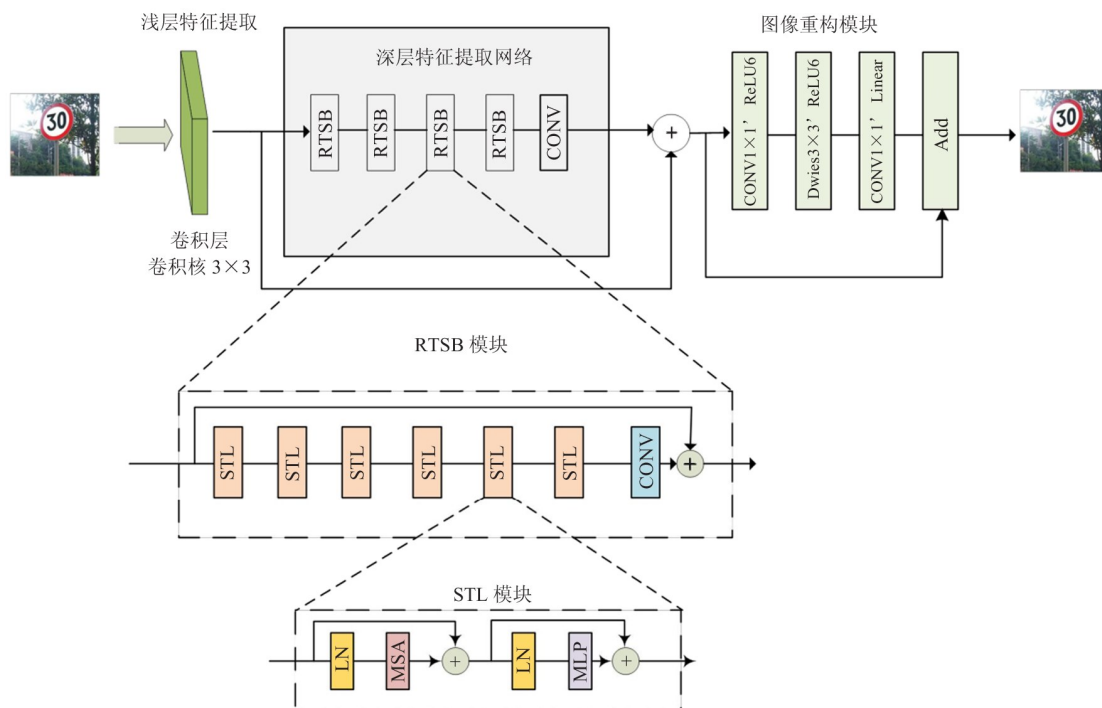


图 3 改进后 SwinIR 网络架构图

分别是图像的高度、宽度和通道数),使用  $3 \times 3$  卷积层  $H_{SF}()$  来提取浅层特征  $F_0 \in R^{H \times W \times C}$ 。

$$F_0 = H_{SF}(I_{LQ}) \quad (6)$$

进一步,从  $F_0$  中提取深层特征  $F_{DF} \in R^{H \times W \times C}$ 。

$$F_{DF} = H_{DF}(F_0) \quad (7)$$

式中:  $H_{DF}()$  是深度特征提取模块,它包含  $k$  个残差 Swin-Transformer 块和 1 个卷积核大小为  $3 \times 3$  的卷积层,并在 Swin-Transformer 层 (Swin-Transformer layer, STL) 中引入滑动窗口注意力机制。如图 3 所示,给定大小为  $H \times W \times C$  的输入,首先,对窗口内的特征图进行层归一化 (layer normalization, LN) 处理<sup>[17]</sup>,即计算该样本在每个特征上的均值和方差。然后,对每个窗口进行多头自注意力机制 (window-based multi-head self-attention, W-MSA) 计算,再将输出的特征经过具有 2 个全连接层构成的多层感知器 (multilayer perceptron, MLP) 进行维度变换,以便下一次的运算处理。整个过程被表述为

$$\hat{z}^l = W - MSA(LN(z^{l-1})) + z^{l-1} \quad (8)$$

$$z^l = MLP(LN(\hat{z}^l)) + \hat{z}^l \quad (9)$$

式中:  $l$  代表 Swin-Transformer 的层数,  $z^{l-1}$  表第  $l-1$  层输入的特征图,  $W - MSA$  为窗口内的多头自注意力机制计算操作,  $LN$  为层归一化,  $MLP$  为多层感知机。

$H_{DF}()$  中间特征层的输出  $F_1, F_2, \dots, F_k$  和深度特征  $F_{DF}$  的公式表示如下:

$$F_i = H_{RSTB_i}(F_{i-1}), i = 1, 2, \dots, k \quad (10)$$

$$F_{DF} = H_{CONV}(F_k) \quad (11)$$

式中:  $H_{RSTB_i}()$  表示第  $i$  个残差 Swin-Transformer 模块 (residual Swin-Transformer block, RSTB),  $H_{CONV}$  是在末尾的卷积层。使用卷积层作为特征提取的最后一层,将卷积运算的归纳偏置引入到 Transformer 中,为后期的浅层特征和深层特征的聚合奠定更好的基础。

在 SwinIR 中的最后一个模块,即图像重建模块,将原有的单一卷积层替换为由 3 个卷积层组成的倒残差结构,并在最后一个卷积层末尾引入线性激活函数,替代传统的 ReLU 激活函数。由于倒残差结构的设计特点是两头细、中间粗,因此该结构末尾卷积层的输出形成了一个低维特征向量,采用

ReLU 激活函数可能导致低维特征信息的大量损失<sup>[18]</sup>,因此,选择使用线性激活函数,以降低在低维特征空间中可能造成的信息损失。具体流程为先用  $1 \times 1$  的卷积核对图像进行升维操作,增大图像的通道数。然后在中间一层引入深度可分离卷积 (depth-wise separable convolution, DWConv)<sup>[19]</sup> 代替普通的卷积操作,该卷积具有更少的参数量和计算量,同时一定程度上保持了模型的表达能力。最后再用卷积核大小为  $1 \times 1$  的卷积层进行降维处理。在图像重建模块中采用倒残差结构可以让网络更好地融合浅层特征  $F_0$  和深层特征  $F_{DF}$ ,有效去除对抗性扰动,并减少模型的计算量,缩短模型的收敛耗时。

### 2.3 图像生成

考虑到一维标签信息与高维图像信息间存在巨大的维度差异,在网络处理不同维度数据时容易出现匹配模糊问题。且高维图像信息比低维文本信息包含更多的细节和特征,通过高维数据丰富的特征表达能力,能够使模型更好地捕捉数据中的复杂关系。由此考虑将图像的文本标签信息转化为图像,用图像类别“一致性”检测取代文本标签“一致性”检测,以解决文本标签与图像特征之间的匹配模糊问题,后续实验验证了其有效性。

CWGAN-GP 是一种生成对抗网络的变体,其结合了条件生成和 Wasserstein GAN 的思想,并加入了梯度惩罚机制<sup>[20]</sup>。CWGAN-GP 可以通过标签信息生成对应类别图像,即实现了将标签信息从文本信息转化为图像的语义信息。CWGAN-GP 结构框图如图 4 所示,其基本工作流程如下所述。

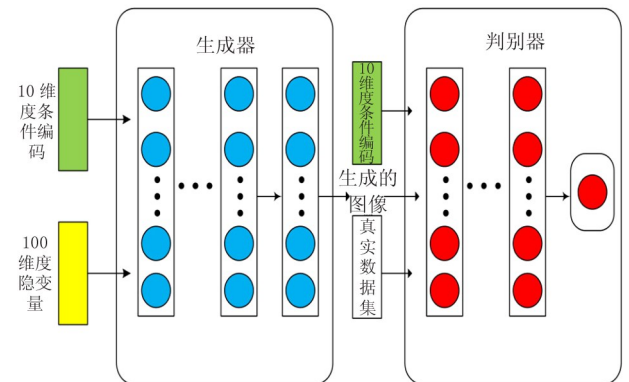


图 4 CWGAN-GP 结构框图

首先,对图像标签按类别数目进行编码,确保每个类别都对应唯一的编码值<sup>[21]</sup>。生成模型负责通过随机噪声和条件信息生成与特定类别和特定属性相关的图像。通过将标签编码作为输入之一,生成模型可以更加精确地控制生成图像的外观和特征。而判别模型只能根据输入的图像和与之相关的条件编码来进行分类或判断。通过结合图像和条件编码进行判断,判别模型可以更好地理解和利用输入图像的特征,提升分类的准确率和鲁棒性。

接着,开始训练模型。依据 Wasserstein GAN 的思想定义生成器和判别器的损失函数,如式(12)和(13)所示。迭代训练生成器和判别器的过程是 CWGAN-GP 网络训练的核心。从真实数据和潜在噪声空间中采样数据,通过交替更新判别器和生成器的参数,判别器尝试最大化真实数据和伪造样本之间的 Wasserstein 距离,同时生成器试图最小化这个距离。在判别器更新时,还要添加梯度惩罚项以确保其 Lipschitz 连续性。这个过程持续进行直至达

到训练停止条件。

$$L(D) = -E_{x \sim P_r}[D(x|y)] + E_{\tilde{x} \sim P_g}[D(\tilde{x}|y)] + \lambda E_{\hat{x} \sim P_{\hat{x}}}[(\|\nabla_{\hat{x}} D(\hat{x}|y)\|_2 - 1)^2] \quad (12)$$

$$L(G) = -E_{\tilde{x} \sim P_g}[D(\tilde{x}|y)] \quad (13)$$

式中:  $y$  代表标签信息,  $x$  代表生成的图像,  $P_r$  是真实数据分布,  $P_g$  是生成器生成数据的分布,  $\tilde{x}$  表示生成的数据,  $\hat{x}$  是生成的数据和真实的数据的线性插值,  $\lambda$  代表梯度惩罚系数。

最后,训练完成后的 CWGAN-GP 网络根据标签信息生成不同类别的图像,用于下一步的分类对比。

## 2.4 图像分类

考虑到计算工作量及检测效率,该模块采用简单的卷积神经网络完成分类任务。设置图像分类部分的目的在于判断重构后的“去对抗化”图像和依据文本标签的“生成”图像类别是否一致。以 ResNet-18 分类网络为例,其网络架构如图 5 所示。

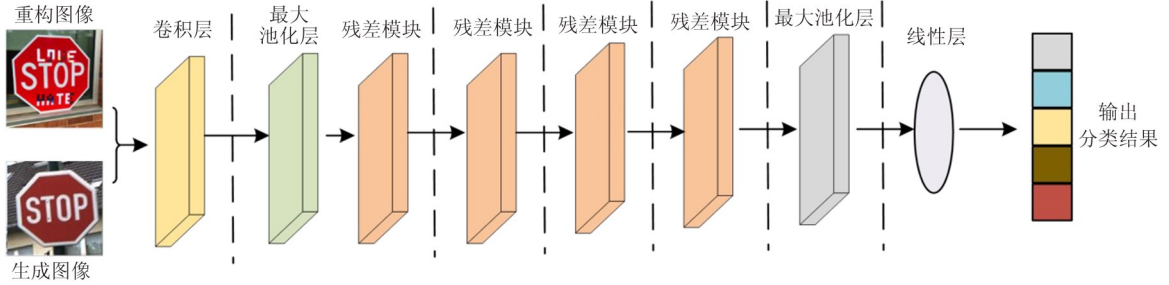


图 5 ResNet-18 网络架构图

将经过 SwinIR 网络去除扰动性噪声后的重构图像与 CWGAN-GP 网络依据标签信息生成的图像,分别输入至 ResNet-18 网络中进行分类,记录并对比二者分类结果。如果分类网络对这 2 种图像都给出相同的类别标签,则将其归类为正常样本;否则,将其判定为对抗样本。

## 3 实验与分析

### 3.1 实验设置

实验平台选用 Win10 操作系统和 PyTorch 网络架构,并采用 CUDA11.3 对训练进行加速。在硬件

设施方面,处理器为 i7-13700KF,显卡为 RTX 4070Ti 12 GB 显存。

实验使用 MNIST<sup>[6]</sup>数据集和 GTSRB<sup>[6]</sup>数据集,并使用 Foolbox<sup>[22]</sup>工具针对不同分类网络生成 FGSM、C&W、BIM、DeepFool 这 4 种类别的对抗样本,每一类别样本数量为 1 000 张。使用 MNIST 和 GTSRB 数据集分别训练 VGG-16<sup>[23]</sup>、ResNet-18<sup>[24]</sup>和 GoogLeNet<sup>[25]</sup>分类神经网络,训练次数为 200,完成后保存模型参数,并记录其准确率,各分类模型准确率如表 1 所示。

在训练和测试阶段,该方法中图像重构模块使用的滑动窗口大小为  $7 \times 7$ ,RTSB 模块和 STL 层的

数量均为 6,注意力头个数为 6,随机失活率(Drop-out)设置为 0.05,学习率设置为 0.000 1,批大小(batch-size)设置为 64,训练轮次为 400。在图像生成阶段,随机噪声的维度设定为 100,batch-size 设置为 64,优化器选用学习率为 0.000 2 的 Adam 优化器,训练轮次为 500。

表 1 训练后各分类模型准确率 %

模型	数据集	
	MNIST	GTSRB
VGG-16	98.47	98.03
ResNet-18	99.41	98.55
GoogLeNet	99.32	98.29

为检验本文设计图像重构模块的有效性,在 GTSRB 数据集和 ResNet-18 分类模型上进行消融实验;同时为验证图像生成模块的必要性,基于 MNIST、GTSRB 数据集和 ResNet-18 分类网络进行了对比实验。

为验证 Recon-Gen 方法的总体性能,与当前性能较优的 Argos<sup>[6]</sup>、局部固有维度(local intrinsic dimensionality, LID)<sup>[26]</sup>、特征压缩(feature squeezing, FS)<sup>[27]</sup>检测方法进行对比实验。相较于其他对抗样本检测方法在训练阶段需要使用与正常样本数量相同的对抗样本,Recon-Gen 方法在训练阶段不需要使用对抗样本进行训练,仅使用正常样本即可。

### 3.2 评价指标

评价指标采用通用的准确率(accuracy, ACC)和 F1 分数(F1-score, F1)这 2 个指标来评价检测性能的好坏。

准确率是用于评估分类模型性能的一种指标,表示模型正确分类的样本数量与总样本数量之比。通常情况下,准确率越高,表示模型在分类任务中的性能越好。该指标可表示为式(14)。

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \quad (14)$$

F1 分数是一种常用的用于评估二分类模型性能的指标,该指标综合考虑了模型的精确率(Precision)和召回率(Recall),使得模型在二者之间取得平衡。通常情况下, F1 分数越高,表示模型在二分

类任务中的性能越好。F1 分数可表示为式(15)。

$$F1 = \frac{2TP}{2TP + FN + FP} \quad (15)$$

在测试阶段使用 1 000 张对抗样本和 1 000 张正常样本来举例说明,分类检测结果的混淆矩阵热力图如图 6 所示。图中,正例代表对抗样本,反例代表正常样本。左上方真正例(true positive, TP)和右下方真反例(true negative, TN)表示真实标签和预测标签一致的样本类别;右上方假反例(false negative, FN)和左下方假正例(false positive, FP)表示真实标签和预测标签不一致的样本类别。在该图中,1 000 张对抗样本有 917 张被成功检测,而 83 张被误判为正常类别;1 000 张正常样本中有 977 张被正常分类,而 23 张被误判为对抗样本。

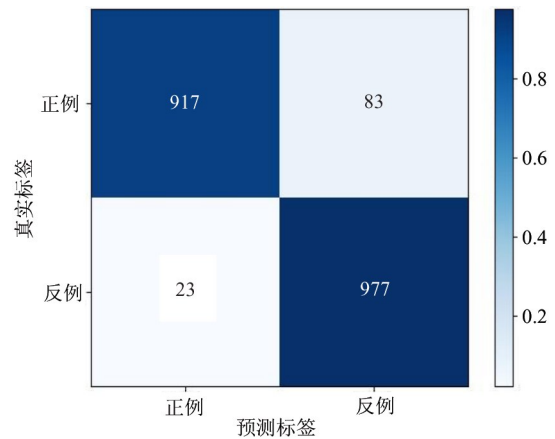


图 6 分类检测结果混淆矩阵热力图

### 3.3 消融实验

为验证图像重构模块在 Recon-Gen 方法中的必要性,本文在 GTSRB 数据集和 ResNet-18 分类模型上进行消融实验。实验包含 3 个部分:(1)无图像重构模块;(2)图像重构模块为初始 SwinIR 网络;(3)图像重构模块为引入倒残差结构和线性激活函数的 SwinIR 网络,实验结果如表 2 所示。

表 2 不同重构方法下检测结果对比 %

方法	评价指标	
	ACC	F1
无重构模块	56	65
SwinIR	73	79
改进后 SwinIR	81	84

由表 2 可知,使用 SwinIR 网络重构输入图像相较于无图像重构模块直接进行检测,其  $ACC$  提升 30.3%,  $F1$  分数提升 21.5%,验证了图像重构模块消除对抗性扰动在检测方法中的必要性。在引入倒残差结构和线性激活函数后的 SwinIR 网络相较于初始的 SwinIR 网络在  $ACC$  指标上提升 10.9%,在  $F1$  分数指标上提升 6.3%,验证了本文所提改进方法的有效性。改进后的 SwinIR 重构网络能够更好地学习图像中的深层语义特征并消除对抗性扰动<sup>[28]</sup>,使得对抗样本经过重构后被分类为与原类别不同的其他类别,而正常样本仍保持原有类别,从而有效检测出对抗样本。

为验证图像生成模块对提升系统检测准确率的帮助,本文基于 MNIST、GTSRB 数据集和 ResNet-18 分类网络进行对比实验。实验包含 2 个部分:(1) 无图像生成模块,直接使用图像原有的标签信息与重构后图像的分类结果对比;(2) 使用 CWGAN-GP 网络通过标签生成对应类别图像,再与重构后的图像输入至同一分类网络中进行分类并对比,实验结果如表 3 所示。

表 3 有无图像生成模块下检测结果对比 %

方法	MNIST		GTSRB	
	$ACC$	$F1$	$ACC$	$F1$
无生成模块	81	84	79	81
有生成模块	83	88	80	83

分析表 3 可知,添加图像生成模块后,在 MNIST 数据集上  $ACC$  提升 2.4%,  $F1$  分数提升 4.7%,在 GTSRB 数据集上  $ACC$  提升 1.2%,  $F1$  分数提升 2.4%。其原因在于:标签信息是一维文本信息,与高

维图像之间存在巨大的维度差异,神经网络不易建立起其中的映射关系,若直接用图像原有标签与重构后图像的分类结果进行对比,可能会存在较大误差,且缺乏对各类识别网络的普适性。将标签信息转化为图像后,可以简化数据处理流程,避免由于输入数据维度不同而造成的潜在错误,且高维图像数据更容易进行可视化操作,便于理解和解释模型的决策过程,也有助于后续的调试和改进模型。

### 3.4 对比实验

#### 3.4.1 准确率和 $F1$ 分数对比实验

将 Recon-Gen 检测算法与 Argos、LID、FS 传统检测方法,基于 MNIST 和 GTSRB 数据集,分别在 ResNet-18、VGG-16、GoogLeNet 这 3 种分类模型上进行验证。

分析表 4 数据可知,在 VGG-16 分类网络上检测 4 类对抗样本时,Recon-Gen 方法相较于其他 3 种方法是性能最优的检测方法,在 MNIST 数据集上其准确率提升了 7.1% ~ 27.1%,  $F1$  分数提升了 3.3% ~ 15.0%;在 GTSRB 数据集上其准确率提升 -1.2% ~ 16.7%,  $F1$  分数提升 -2.3% ~ 18.5%。在 GTSRB 数据集上检测 BIM 对抗样本时,Recon-Gen 方法的准确率和  $F1$  分数略低于 Argos。分析其原因:首先,BIM 沿损失函数梯度上升的方向多次添加扰动,导致图像失真严重,在图像重构阶段无法完全还原其深层语义结构特征,从而无法有效对图像进行“去对抗化”操作,进而影响到最终的检测结果。其次,VGG-16 分类网络使用小尺寸卷积核和较小的池化窗口可能无法捕捉到更大范围的空间局部性特征,且 VGG-16 网络深度不足,仅由 13 层卷积层和 3 层全连接层组成,在输入数据失真严重时,容易输出

表 4 各检测方法在 VGG-16 分类网络上的指标对比 %

攻击类型	Recon-Gen				Argos				LID				FS			
	MNIST		GTSRB		MNIST		GTSRB		MNIST		GTSRB		MNIST		GTSRB	
	$ACC$	$F1$	$ACC$	$F1$	$ACC$	$F1$	$ACC$	$F1$	$ACC$	$F1$	$ACC$	$F1$	$ACC$	$F1$	$ACC$	$F1$
FGSM	90	94	89	93	84	91	87	92	60	67	81	88	66	78	76	86
C&W	95	97	82	89	68	80	71	81	66	78	74	80	82	90	68	78
DeepFool	87	94	91	96	66	77	73	80	58	66	74	80	79	83	78	81
BIM	89	92	82	84	70	80	83	86	48	55	73	81	47	55	71	80

错误的分类结果。Argos 方法则是基于视图生成机制,将输入图像划分为几个区域,基于不同部分重构图像,通过比较多个视图之间的预测结果来检测对抗性样本。对于 BIM 攻击而言,其并未改变图像中所有像素点,即使某些视图被攻击者成功欺骗,其他视图仍然可以保持正确的预测。

分析表 5 数据可知,在 ResNet-18 分类网络上检测 4 类对抗样本时,Recon-Gen 方法相较于其他检测方法中的最优检测方法在 MNIST 数据集上

的准确率提升 0.0% ~ 24.0%, F1 分数提升 2.2% ~ 11.9%;在 GTSRB 数据集上检测准确率提升 9.3% ~ 21.1%, F1 分数提升 6.7% ~ 13.4%。由于 ResNet-18 引入残差学习的概念,利用残差块学习原始输入与期望输出之间的差异,且 ResNet-18 相较于传统的深度卷积神经网络,具有更深的层级结构,能够捕捉更丰富和抽象的特征,故而在面对不同对抗攻击时,ResNet-18 可以较好地保留图像深层语义特征,有助于 Recon-Gen 方法实现对抗样本检测。

表 5 各检测方法在 ResNet-18 分类网络上的指标对比

%

攻击类型	Recon-Gen				Argos				LID				FS			
	MNIST		GTSRB		MNIST		GTSRB		MNIST		GTSRB		MNIST		GTSRB	
	ACC	F1	ACC	F1	ACC	F1	ACC	F1	ACC	F1	ACC	F1	ACC	F1	ACC	F1
FGSM	81	90	89	93	75	82	78	82	71	80	68	80	65	72	68	78
C&W	90	94	82	89	88	92	75	83	75	83	66	77	70	82	45	61
Deepfool	93	94	92	96	67	77	56	64	75	84	71	80	72	83	77	90
BIM	78	85	86	92	78	82	65	77	66	76	63	74	56	72	71	83

分析表 6 数据可知,在 GoogLeNet 分类网络上检测 4 类对抗样本时,Recon-Gen 方法在 MNIST 数据集和 GTSRB 数据集上的准确率和 F1 得分 2 项指标均优于其他 3 种检测方法,在 MNIST 数据集上的准确率提升了 5.2% ~ 21.6%, F1 分数提升了 4.4% ~ 11.8%;在 GTSRB 数据集上准确率提升了

8.1% ~ 26.7%, F1 分数提升了 6.10% ~ 11.25%。由于 GoogLeNet 引入了 Inception 模块,该模块使用不同大小的卷积核捕捉不同尺度的特征,从而更好地捕获输入数据中的多尺度信息,保留较大跨度范围内的图像语义特征,利于 Recon-Gen 方法实现对抗样本检测。

表 6 各检测方法在 GoogLeNet 分类网络上的指标对比

%

攻击类型	Recon-Gen				Argos				LID				FS			
	MNIST		GTSRB		MNIST		GTSRB		MNIST		GTSRB		MNIST		GTSRB	
	ACC	F1	ACC	F1	ACC	F1	ACC	F1	ACC	F1	ACC	F1	ACC	F1	ACC	F1
FGSM	80	91	87	92	76	82	80	84	72	80	70	80	66	71	70	82
C&W	92	94	80	89	84	90	74	80	72	79	65	75	68	77	50	64
Deepfool	90	92	90	93	67	75	59	65	74	82	71	76	73	85	68	84
BIM	75	85	83	87	67	76	67	75	65	73	60	72	55	68	72	82

综合来看,Recon-Gen 方法相比其他 3 种检测方法在 DeepFool 攻击上的检测率提升最为显著。原因在于 DeepFool 是一种迭代的、渐进的攻击方式,生成的对抗性扰动非常小<sup>[29]</sup>。对于 Argos、LID、FS 等检测方法不易检测出细微的对抗性扰动,而对于 Recon-Gen 检测方法而言检测过程不依赖于图像

某一个区域像素值的变化,而是通过图像重构模块学习图像中的语义信息进而完成检测,因此可以检测出包含细微扰动的对抗样本。在 BIM 攻击算法的检测中,4 种方法表现均不理想,其原因在于 BIM 攻击采用迭代的方式逐步调整对抗样本,每次迭代都根据模型的梯度信息来更新样本,使其更接近能

引发模型误判的边界<sup>[30]</sup>。这种迭代过程可以确保对抗样本在保持微小扰动的同时,最大化地改变模型的输出,从而使其难以被检测出来;对于 Recon-Gen 检测方法而言,BIM 通过多次攻击修改输入图像的像素点,使得图像失真严重,经过 SwinIR 重构图像后仍无法还原其深层语义特征,导致检测效果不理想。

### 3.4.2 收敛耗时对比实验

收敛耗时是對抗样本检测方法性能评估的重要指标。将本文所提检测方法与当前性能较优的 3 种检测方法均训练至 80% 检测准确率,记录所需的时长,各方法耗时对比如图 7 所示。

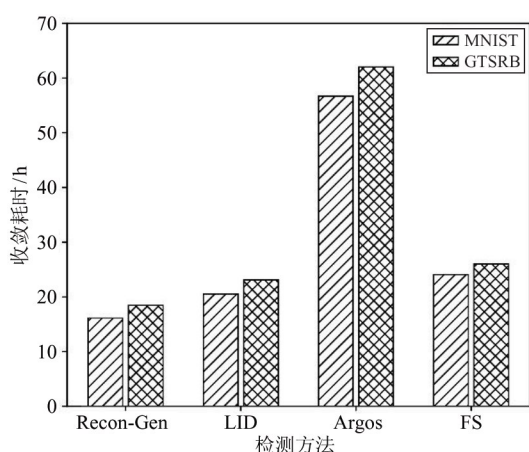


图 7 4 种检测方法在 2 种数据集上收敛耗时对比

Recon-Gen 方法在 MNIST 数据集上训练至 80% 准确率所需的收敛耗时为 18.45 h,而在 GTSRB 数据集上的耗时为 16.12 h,相较于 LID、Argos、FS 方法,其收敛耗时缩短了 25% ~ 78%。这主要源于 Recon-Gen 方法的独特设计,该方法无需对抗样本参与训练,也无需先验地了解被攻击模型的结构和参数。通过条件式生成对抗网络依据标签信息生成完整图像,避免了繁琐的模型训练过程,从而在收敛耗时性能上取得优势。相比之下,FS 算法需要对输入数据进行多次变换和处理,且需要对检测网络中每一层的输出进行特征统计,这会增加算法的计算复杂度。LID 算法中存在一些参数,如邻域大小等,对算法的性能有很大影响。选择合适的参数值需要大量的试验和调优,增加了训练的时间成本。而 Argos 方法则需要依次生成图像的每个像素点,且 FS、LID、Ar-

gos 方法均需要大量对抗样本参与训练,导致训练过程的收敛耗时较长。本文方法最终将对抗样本检测问题转化为分类问题,分类问题算法的复杂度低,分类神经网络检出结果的速率快,相较于其他检测方法针对不同输入图像进行特征差异检测,该方法检出对抗样本所需时间更少,实时性更强。综上所述,Recon-Gen 方法在检测实时性和训练时效上体现出了较大的优势。

## 4 结论

针对现有对抗样本检测方法存在的准确率低且模型训练收敛速度慢的问题,本文提出了一种联合图像重构和图像生成技术的对抗样本检测方法。首先,使用改进后的 SwinIR 图像重构网络,去除对抗样本中的对抗性扰动;接着,采用条件生成式对抗网络,通过标签信息生成图像;最后,将重构图像与生成图像送至识别网络中进行分类,通过比较分类结果的一致性判断是否为对抗样本。实验结果表明,本文所提检测方法对 FGSM、BIM、C&W、DeepFool 攻击生成的对抗样本具有较好的检出性能,且该方法的收敛速度较快,在训练时效上同样体现出优势。

未来的研究将尝试更有效的图像重构网络,在尽可能保留图像深层语义特征的同时,有效去除对抗性扰动;另外在图像生成方面,尝试生成质量更高的生成式对抗网络,用以提升对抗样本检测的准确性。同时在网络规模等方面进行积极探索,减少网络规模及运行所需的计算资源。

### 参考文献

- [1] 宋留静, 赵泽方, 马宇翔, 等. 基于多维语义特征与层次注意力机制的讽刺识别[J]. 高技术通讯, 2024, 34(5): 453-462.
- [2] MIAO Y, CHEN C, PAN L, et al. Machine learning-based cyber attacks targeting on controlled information: a survey[J]. ACM Computing Surveys, 2021, 54(7): 1-36.
- [3] 牛京玉, 胡瑜, 李玮, 等. 基于持续强化学习的自动驾驶赛车决策算法研究[J]. 高技术通讯, 2024, 34(1): 1-14.
- [4] YANG H Y, ZHANG Z X, XIE L X, et al. Network se-

- curity situation assessment with network attack behavior classification[J]. *International Journal of Intelligent Systems*, 2022,37(10):6909-6927.
- [ 5 ] SZEGEDY C, ZAREMBA W, SUTSKEVER I, et al. Intriguing properties of neural networks[EB/OL]. (2014-02-19) [2024-07-15]. <http://arxiv.org/pdf/1312.6199>.
- [ 6 ] KIANI S, AWAN S, LAN C, et al. Two souls in an adversarial image: towards universal adversarial example detection using multi-view inconsistency: annual computer security applications conference [EB/OL]. (2021-10-11) [2024-07-15]. <http://arxiv.org/pdf/2109.12459>.
- [ 7 ] 杨宏宇, 杨帆. 基于图像去噪和图像生成的对抗样本检测方法[J]. *湖南大学学报(自然科学版)*, 2023(8):72-81.
- [ 8 ] XU W, EVANS D, QI Y. Feature squeezing: detecting adversarial examples in deep neural networks[EB/OL]. (2017-12-05) [2024-07-15]. <http://arxiv.org/pdf/1704.01155>.
- [ 9 ] METZEN J H, GENEWEIN T, FISCHER V, et al. On detecting adversarial perturbations[EB/OL]. (2017-02-21) [2024-07-15]. <http://arxiv.org/pdf/1702.04267>.
- [ 10 ] LIANG J, CAO J, SUN G, et al. SwinIR: image restoration using swin transformer[C]//2021 IEEE/CVF International Conference on Computer Vision Workshops. Montreal, Canada: IEEE, 2021:1833-1844.
- [ 11 ] ZHENG M, LI T, ZHU R, et al. Conditional Wasserstein generative adversarial network-gradient penalty-based approach to alleviating imbalanced data classification[J]. *Information Sciences*, 2020, 512:1009-1023.
- [ 12 ] GOODFELLOW I J, SHLENS J, SZEGEDY C. Explaining and harnessing adversarial examples [EB/OL]. (2014-12-20) [2024-07-15]. <http://arxiv.org/pdf/1412.6572v1>.
- [ 13 ] KURAKIN A, GOODFELLOW I, BENGIO S. Adversarial machine learning at scale [EB/OL]. (2017-02-11) [2024-07-15]. <http://arxiv.org/pdf/1611.01236>.
- [ 14 ] CARLINI N, WAGNER D. Towards evaluating the robustness of neural networks[C]//2017 IEEE Symposium on Security and Privacy. San Jose, USA: IEEE, 2017:39-57.
- [ 15 ] MOOSAVI-DEZFOOLI S, FWAZI A, FROSSARD P. Deepfool: a simple and accurate method to fool deep neural networks [C] //2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE, 2016:2574-2582.
- [ 16 ] LIU Z, LIN Y, CAO Y, et al. Swin Transformer: hierarchical vision transformer using shifted windows [C] //2021 IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE, 2021:9992-10002.
- [ 17 ] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, USA: ACM, 2017:6000-6010.
- [ 18 ] SANDLER M, HOWARD A, ZHU M, et al. MobileNetV2: inverted residuals and linear bottlenecks [C] //2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE, 2018:4510-4520.
- [ 19 ] CHOLLET F. Xception: deep learning with depthwise separable convolutions [C] //2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE, 2017:1800-1807.
- [ 20 ] 李阳, 蒋三新. 基于改进生成对抗网络的无监督晶圆缺陷检测[J]. *电子测量技术*, 2023,46(6):91-99.
- [ 21 ] GAUTHIER J. Conditional generative adversarial nets for convolutional face generation[EB/OL]. [2024-07-15]. [https://cs231n.stanford.edu/reports/2015/pdfs/jgauthier\\_final\\_report.pdf](https://cs231n.stanford.edu/reports/2015/pdfs/jgauthier_final_report.pdf).
- [ 22 ] ESMAEILPOUR M, CARDINAL P, KOERICHA L. From environmental sound representation to robustness of 2D CNN models against adversarial attacks [EB/OL]. (2021-01-18) [2024-07-15]. <http://arxiv.org/pdf/2007.13703v3>.
- [ 23 ] XIAO T, ZHANG J, YANG K, et al. Error-driven incremental learning in deep convolutional neural network for large-scale image classification [C] // Proceedings of the 22nd ACM International Conference on Multimedia. Orlando, USA: ACM, 2014:177-186.
- [ 24 ] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition [C] //2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE, 2016:770-778.
- [ 25 ] SZEGEDY C, LIU W, JIA Y, et al. Going deeper with convolutions [C] //2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA: IEEE,

- 2015:1-9.
- [26] MA X, LI B, WANG Y, et al. Characterizing adversarial subspaces using local intrinsic dimensionality[EB/OL]. (2018-03-12) [2024-07-15]. <http://arxiv.org/pdf/1801.02613v2>.
- [27] XU W, EVANS D, QI Y. Feature squeezing: detecting adversarial examples in deep neural networks[EB/OL]. (2017-12-05) [2024-07-15]. <http://arxiv.org/pdf/1704.01155>.
- [28] ZHOU S, ZHU T, YE D, et al. Boosting model inversion attacks with adversarial examples[J]. IEEE Transactions on Dependable and Secure Computing, 2024, 21 (3): 1451-1468.
- [29] BALAADITYA M, DUNSTON S D. Analysis of the effect of adversarial training in defending efficientnet-B0 model from deepfool attack[C]//2023 3rd International Conference on Intelligent Communication and Computational Techniques. Jaipur, India: IEEE, 2023:1-7.
- [30] BANIECKI H, BIECEK P. Adversarial attacks and defenses in explainable artificial intelligence: a survey[EB/OL]. (2024-02-13) [2024-07-15]. <http://arxiv.org/pdf/1704.01155>.

## Adversarial example detection method based on image generation and image reconstruction technology

LI Baoping, XIA Yuhao

(College of Physics and Electronic Information, Henan Polytechnic University, Jiaozuo 454000)

### Abstract

Adversarial example attack is one of the main security threats to recognition networks. In order to solve the problems of low detection accuracy caused by blurred classification boundaries and slow training convergence rates resulting from the participation of a large number of adversarial samples, a method of adversarial example detection based on image reconstruction and image generation technologies is proposed. Firstly, an image reconstruction network implemented by convolutional layer and Swin-Transformer is designed to restore the semantic information and remove the adversarial noise. Then, a conditional generative adversarial network is used to generate images according to image classification label information. Finally, the reconstructed and generated images are input into the convolutional recognition network for classification, and the consistency of classification result is used to determine whether the input image is an adversarial example. The detection method converts the adversarial sample detection problem into an image classification problem, it does not need adversarial samples to participate in model training, and it is not necessary to know the attacker's attack types, the attacked model's structure and parameters in advance. Experimental results on VGG-16, ResNet-18, GoogLeNet classification network, MNIST and GTSRB data sets show that compared with other classical detection methods, the average recognition accuracy of this detection method is increased by 4.75% - 22.86%, *F1* score is increased by 3.40% - 13.64%, which proves its superiority.

**Key words:** adversarial example, image classification, Swin-Transformer, image reconstruction, convolutional neural network (CNN)