

DNN系统高风险对抗测试用例生成方法研究

贾云飞, 张寓涵, 梁文龙, 吴仁彪

(天津市智能信号与图像处理重点实验室, 天津 300000)

摘要: 以深度神经网络(DNN)为核心的软件系统在高风险场景下的安全性测试能力尤为重要, 严重误分类可能引发重大故障。为此, 基于SCSM矩阵, 通过引入高斯核函数来平滑类别差异, 并通过调节参数控制错误严重等级; 同时利用置信度加权严重性指标, 并结合逐层反推筛选关键像素添加扰动, 提升对抗样本的不可感知性。在多个数据集上的实验结果表明, 在不可感知性指标相近时, 所提方法生成测试用例的ADM值优于基线方法, 尤其是在STL-10数据集上的ADM提升约23%, 表明其触发严重错误的能力更强。该方法可根据测试需求生成不同风险等级的用例, 从而有效揭示DNN模型的深层缺陷, 为安全关键系统测试提供精准可靠的数据支持。

关键词: 软件测试; 对抗攻击; 测试用例生成; 深度神经网络

DOI: 10.11907/rjdk.251627

扫描二维码阅读全文:



中图分类号: TP183

文献标识码: A

文章编号: 1672-7800(2025)012-0085-06

Research on High-Risk Adversarial Test Case Generation Method for DNN Systems

JIA Yunfei, ZHANG Yuhan, LIANG Wenlong, WU Renbiao

(Tianjin Key Laboratory of Intelligent Signal and Image Processing, Tianjin 300000, China)

Abstract: The security testing capability of software systems with deep neural networks (DNNs) as the core is particularly important in high-risk scenarios, as severe misclassification may lead to major failures. To this end, based on the SCSM matrix, Gaussian kernel functions are introduced to smooth out class differences, and error severity levels are controlled by adjusting parameters; At the same time, using confidence weighted severity indicators and combining layer by layer backpropagation to filter key pixels and add perturbations to enhance the imperceptibility of adversarial samples. The experimental results on multiple datasets show that when the imperceptibility indicators are similar, the proposed method generates better ADM values for test cases than the baseline method, especially with an ADM improvement of about 23% on STL-10 dataset, indicating a stronger ability to trigger serious errors. This method can generate test cases with different risk levels based on testing requirements, effectively revealing the deep flaws of the DNN model and providing accurate and reliable data support for safety critical system testing.

Key Words: software testing; adversarial attack; test case generation; deep neural network

0 引言

深度神经网络(Deep Neural Network, DNN)作为人工智能的核心技术, 已在医疗预测、入侵检测和自动驾驶等高安全要求领域广泛应用^[1-3]。然而, 在这些高风险场景中, DNN模型的误判可能会引发严重后果。以自动驾驶系统为例, 环境干扰导致的错误识别, 如“幻影刹车”或未能识别过马路行人已暴露出DNN在真实环境中的脆弱

性^[4-5]。尽管现有系统普遍通过了覆盖率测试并具备一定的容错机制, 但是一旦发生严重误分类, 仍可能直接引发重大事故。

近年来, 越来越多的研究指出, 对抗严重性, 即错误导致的后果等级, 应成为评估DNN安全性的关键维度^[6]。现有测试方法在错误严重性控制和对抗样本生成精度方面仍存在明显不足^[7]。因此, 开展面向高风险对抗测试用例(指可导致DNN系统严重误判的测试用例)的生成方法研究, 对于提升DNN系统的可靠性和安全性具有重要的理

收稿日期: 2025-09-06

基金项目: 国家重点研发计划项目(2022YFB3904300); 某型机械设备智能制造系统开发项目(H01420210199)

作者简介: 贾云飞(1979-), 男, 天津市智能信号与图像处理重点实验室副教授, 研究方向为软件测评与性能优化、民航运输大数据技术、智能制造技术。

论价值与现实意义。

1 相关研究

DNN的对抗攻击与安全性测试已成为近年来的研究热点。这类研究不仅揭示了DNN模型在鲁棒性上的脆弱性,而且为模型安全性评估提供了重要思路^[8]。

对抗攻击是指通过对输入样本添加人眼难以察觉的微小扰动,使模型产生错误输出的行为^[9]。目前大多数对抗测试方法主要关注如何提高对抗样本的生成效率或攻击成功率,例如通过优化扰动生成算法以提升对抗样本的误导能力^[10-11]。然而,这类方法普遍忽略了一个关键问题:不同的错误分类在实际应用中所导致的后果严重程度可能存在显著差异。

部分学者开始注意到错误严重性的重要性。例如, Sun等^[12]发现不同对抗算法会导致模型置信度出现不同程度的下降,间接反映了模型对特定错误的确信程度。类似地, Ogonna等^[13]探讨了DNN中不确定性与对抗攻击之间的关系,指出置信度与模型脆弱性存在内在关联。这些研究为从错误置信水平角度衡量模型风险提供了初步依据。Filus等^[7]首次明确提出对抗严重性的量化思路,并构建了不相似度度量(Dissimilarity Metric, DM),用于评估对抗样本所触发错误与原始类别的语义距离,从而区分最优与最差故障场景。该方法为生成不同风险等级的对抗样本奠定了基础,标志着测试焦点从“是否错误”转向“错误的严重程度”。

在以上研究基础上,本文提出FGHD-adv(Fast Gradient Sign Highly Dissimilar-Adversarial)算法,通过计算类别

之间的相似度和前向梯度来选择关键像素并生成难以察觉的对抗样本。与现有研究相比,该算法在高风险对抗测试用例生成方法上实现了两项关键改进:一是创新性地引入高斯核函数对SCSM(Service Manager Configuration Manager)矩阵中的类别差异进行平滑处理,通过连续可调的参数 λ 控制错误严重性等级,使生成的测试用例能灵活覆盖从低风险到高风险的各类错误场景,显著增强了严重性控制的精细度和适应性;二是结合前向梯度逐层反推策略,精准识别并仅对贡献度高的关键像素添加扰动,在有效维持对抗样本视觉不可感知性的同时更精准地引导模型产生严重误分类,从而在相似扰动幅度下更有效地揭示DNN的深层缺陷。

2 FGHD-adv算法

作为一种融合类别差异感知与像素级贡献度分析的新型对抗样本生成算法,首先,FGHD-adv运行DNN模型以获得各个分类结果的置信度,并逐层计算前向梯度可反推每个像素对某一分类的贡献度;其次,利用文献[13]中的SCSM矩阵初步得出标签间的相似度排序。现有研究将排列序号作为错误严重程度,为进一步准确评估每一像素对抗结果的贡献,利用核函数加权各个错误的序号。核函数的选取依据为只要能够实现各个错误结果的加权平滑即可,具体参数可根据具体测试场景确定。本文利用高斯核函数平滑排序结果,得出类别差异度指标。综合以上两步,即可计算出每一像素对类别差异度的贡献。整体流程如图1所示。

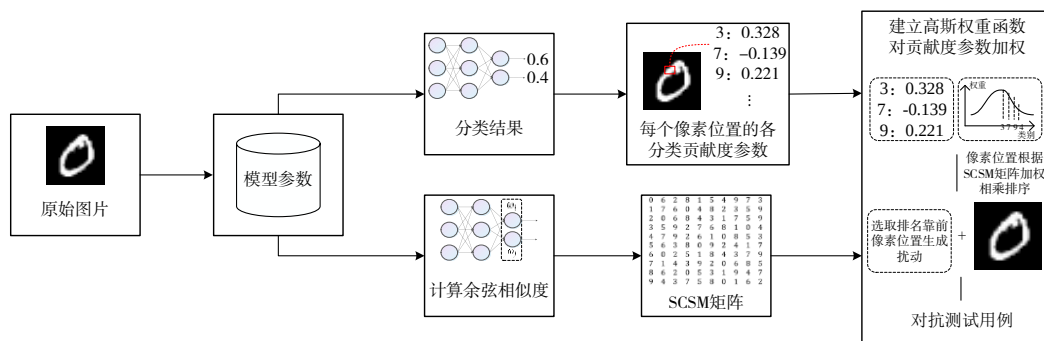


Fig. 1 Flow of the FGHD-adv algorithm

图1 FGHD-adv算法流程

图1所示的生成流程可转化为具体的算法。算法输入包括原始图像 X 、每幅图像的总像素数量 N 、训练好的深度前馈神经网络 F 、分类类别总数 M 、扰动像素选择比例 τ 和扰动量度程度参数 ε ;输出为对抗测试用例 Y 。算法主要步骤为:

(1)对原始样本进行前向传递,得到分类概率,计算每个类在每个像素位置的前向梯度,得到一个 $M \times N$ 的矩阵,表示某个类别的不同像素位置对其贡献度的大小。

(2)利用模型最终分类层的SoftMax激活输出,计算每

个类别代表的神经元与上一层连接的神经元权重向量之间的余弦相似度。按照由大到小的顺序将类别结果对应到原始类别所在每一行的后面,得到SCSM矩阵,第一列代表原始类别,各行中原始类别与其他类别的相似度距离表示差异性。换言之,对抗测试用例错误分类距离原始类别越远,越能揭示严重程度更高的漏洞。

(3)将类别按风险程度从大到小排序,映射到高斯核的 x 轴正半轴上。风险程度最大的类别位于 x 轴的中心,原始类别位于 x 轴最远处。根据高斯函数加权各类别作为

差异度,通过调整参数 λ 可以控制高斯核曲线的形状,使其更平滑或更陡峭。

(4)对每个像素的差异度进行排序,选出前 $\tau\%$ 比例的像素并添加固定值的扰动,即生成对抗测试用例。

FGHD-adv 算法的伪代码为:

输入: 原始输入图片集合 X ; 一张图片的像素总数 N ; 深度神经网络模型 F ; 分类类别总数 M ; 关键像素选择比例 τ ; 扰动幅度参数 ε

输出: 生成的对抗测试用例 Y ;

for data in X

$$J_F(X) = \frac{\partial F(X)}{\partial X} = \left[\frac{\partial F_M(X)}{\partial x_N} \right]_{M \in 0..m, N \in 0..n}; // \text{求解前向梯度 } J_F(X)$$

$Q_n = \text{Contribution Matrix}(J_F(X), N)$; // 组成贡献度矩阵 Q_n

$$CS_{m_0, m_i} = CS(m_0, m_i) = \frac{\omega_{m_0}^T \omega_{m_i}}{\|\omega_{m_0}\| \|\omega_{m_i}\|}; // \text{原始类别 } m_0 \text{ 与其他}$$

类别 m_i 之间的相似度大小

$SCSM = [\text{Sort}(CS[i]) \text{ for } i \text{ in range}(M)]$; // 得到描述类别距离的 SCSM 矩阵

for at each pixel position M of Q_n do

$$K_{\lambda, i}(m_0, m_i) = \frac{1}{\lambda} \exp \left[-\frac{m_i - m_0}{2\lambda} \right] K_{\lambda}; // \text{根据 SCSM 矩阵}$$

类别对应到高斯函数上

$$C_n = \sum_{i=0}^n K_{\lambda, i}(m_0, m_i) \cdot \max(0, Q_n); // \text{根据“混淆度”大小对 } Q_n \text{ 进行赋值加权}$$

if $Q_n < 0$ then; // 只选择对分类正影响的贡献度参数进行加权

: $Q_n \leftarrow 0$;

end if

end for

Sort all elements in C_n in descending order; // 对不同像素位置

加权贡献度后降序排序

$X^* \leftarrow \{Q_n | n \in \text{indices_of_top_}\tau(C_n)\}$; // 选择比例为 τ 的关键像素位置

$$Y = X^* + \varepsilon$$

return Y

end

3 实验方法与结果分析

3.1 实验环境与参数设置

本文实验在 Windows 10 操作系统上完成,处理器为 Intel(R) Core(TM) i5-8400 2.80GHz, GPU 为 NVIDIA GeForce GTX 1050 Ti,系统内存为 16 GB。使用 1.10.1 版本的 PyTorch 作为深度学习框架,开发环境为 Python 3.8, CUDA 版本为 10.2。网络训练超参数为:150 epoch、初始学习率 3×10^{-4} 、AdamW 优化器、cosine 退火至 1×10^{-6} 、batch size 32、交叉熵损失,并辅以随机裁剪、水平翻转及 ImageNet 统计归一化。随机种子固定为 42,确保结果可复现。训练过程中使用早停机制,若验证集准确率连续 20 轮无提升则终止训练并保存最优模型。

3.2 实验模型与数据集

本文在 MNIST 数据集^[14]上对 LeNet-5 模型^[15]进行训练、在 CIFAR-10 数据集^[16]上对 VGG16 模型^[17]进行训练,在 STL-10 数据集^[18]上对 MobileViT 模型^[19]进行训练。表 1 详细描述了本文使用的模型和数据集,并给出了训练后的准确度。

使用 3 个数据集分别训练 3 个模型作为测试对象。由于原始测试集图片数量过多,从每个类别中随机抽取 100 张,共 1 000 张图片,作为待生成对抗测试用例的原始图片。

Table 1 Experimental models and datasets

表 1 实验模型和数据集

数据集	数据集描述	模型	模型描述	训练准确度/%
MNIST	一个包含 70 000 张 28×28 手写数字灰度图像的数据集,涵盖 0~9 的 10 个类别	LeNet-5	卷积神经网络最著名和具有开创性的系统之一,在手写数字识别和图像分类方面取得了显著成功	99.04
CIFAR-10	一个包含 60 000 张 32×32 彩色图像的数据集,包含 10 类日常生活中的常见物体	VGG16	包含 13 个卷积层和 3 个全连接层,是计算机视觉和深度学习领域中最具影响力和最广泛应用的模型之一	91.36
STL-10	该数据集的类别与 CIFAR-10 类似,但图像来源于 ImageNet,共包含 113 000 张 96×96 分辨率的彩色图像	MobileViT	CNN 与 Transformer 的混合架构模型,在多种移动视觉任务中实现最先进的性能,包括分类和检测	87.39

本文实验尝试性探索最优关键像素比例。具体而言,首先固定扰动值,选取前 10%、20%、30%、40%、50%、60% 进行扰动,然后观测对抗错误率,如达标则认为最优。在对 MNIST 数据集、CIFAR-10 数据集、STL-10 数据集图片进行尝试性探索后,3 个数据集的最优关键像素比例分别确定为 0.550、0.400、0.175,后续实验将在固定对应扰动值的基础上进行。第一组实验将 FGHD-adv 算法与当前的对抗生成方法进行有效率比较。第二组实验探究生成测试

用例的质量及有效性,评价核函数选择不同的平滑参数时,所生成的测试用例集触发严重故障的能力。

3.3 评价指标

一个好的对抗测试用例需要由对抗攻击效果、扰动噪声幅度、效率 3 个方面进行综合评价^[20]。本文除了考虑最基本的错误率外,还重点考虑了衡量风险大小的损害度量指标(Damage Metric, DM),用于表示某批测试用例触发严重错误的能力。DM 值的计算步骤为:①将生成后的用例

集输入 DNN 模型,基于余弦相似度构建类别间语义相似度矩阵 SCSSM,并按相似度降序排列每行;②对于每张图片,采用其原始类别对应行查询对抗攻击后的预测类别在该行的索引位置,索引越靠后表示语义偏离越大;③对所有样本的索引取平均,即得到 DM 值。将 DM 值用于模型安全性测试时,可将风险程度评估值定义为:

$$\begin{cases} 0 & \text{原始类别} \\ 0 < DM \leq 0.5 & \text{低风险错误} \\ 0.5 < DM \leq 1 & \text{高风险错误} \\ 1 & \text{最严重错误} \end{cases} \quad (1)$$

当使用 DM 值评测用例集揭露严重错误的能力时,还应考虑用例集触发错误类别的置信度 P_{adv} 。文献[13]和[20]中指出,如果对抗测试用例能够显著降低模型对原始类别的置信度,或提高模型对错误类别的置信度,则说明模型对这些对抗测试用例具有较高的脆弱性。模型在错误分类上的高置信度表明其对抗扰动非常敏感,揭示了模型的潜在脆弱性。因此,DM 指标经过加权错误类别分类置信度后可表示为:

$$ADM = \frac{1}{n} \sum_{i=1}^n (x_i^a, x_i) P_{adv} \quad (2)$$

式(2)考虑了对抗后测试用例 x_i^a 与对抗前测试用例 x_i 在标签上的差异,以及触发错误类别的置信度 P_{adv} ,并将标签与置信度加权。置信度表明 DNN 模型对这一结果的信赖程度。因此,ADM 表达了 DNN 模型对这一错误结果的确信度。

本文使用两个指标综合评价扰动的不可感知性。第一个指标为文献[21]中提到的 ALD_p (平均 L_p 失真),通常用于评价通过在 L_p 范数的基础上添加噪声来生成对抗测试用例的方法。该指标通过计算 L_p 范数 $\| \cdot \|_p$ 来量化测试用例对抗前后的相对误差,用于评估生成的对抗测试用例的视觉不可察觉性,进而表征两张图像之间的差异。计算公式为:

$$ALD_p = \frac{1}{n} \sum_{i=1}^n \frac{\|x_i^a - x_i\|_p}{\|x_i\|_p} \quad (3)$$

式中: P 的取值在 $0 \sim \infty$ 之间; n 为成功生成的对抗样本数量。

第二个指标为文献[22]中的结构相似指标(Structural Similarity Index, SSIM),其通过比较亮度、对比度和结构来衡量两张图像之间的相似度,是衡量人类感知的有效方法。计算公式为:

$$SSIM(x_i^a, x_i) = \frac{(2\mu_{x_i^a} \mu_{x_i} + C_1)(2\sigma_{x_i^a x_i} + C_2)}{(\mu_{x_i^a}^2 + \mu_{x_i}^2 + C_1)(\sigma_{x_i^a}^2 + \sigma_{x_i}^2 + C_2)} \quad (4)$$

式中: μ 为图像块的像素平均值; σ^2 为图像块的像素方差; $\sigma_{x_i^a x_i}$ 为测试用例对抗前后的像素协方差; C_1 、 C_2 为两个常数,用于避免分母为0。

ASS(Average Structural Similarity)表示数据集中多张原始样本与其对应的对抗测试用例的平均结构相似指标。

计算公式为:

$$ASS = \frac{1}{n} \sum_{i=1}^n SSIM(x_i^a, x_i) \quad (5)$$

为采用一个指标直接衡量不可感知性的大小,本文将衡量测试用例与原始样本相似度的指标 ALD_p 和 ASS 分别归一化为 ALD_{PN} 和 ASS_N ,然后加权为一个指标 Imperceptibility 来衡量不可感知性的好坏。这两个指标有相同的重要性,将加权参数均设置为0.5。Imperceptibility 的计算公式为:

$$Imperceptibility = 0.5(1 - ALD_{PN}) + 0.5ASS_N \quad (6)$$

3.4 实验结果与分析

3.4.1 测试用例生成效率比较

在评估对抗攻击方法所生成对抗样本质量的同时,生成效率问题同样不可忽视。具体而言,若某一攻击方法在依赖大量原始样本的前提下仅能生成少量对抗样本,即便其生成的样本在质量上表现优异,亦难以被视为一种高效的攻击策略。由于对抗攻击是否成功的核心判据在于对抗样本是否导致模型输出标签相对于原始类别发生显著偏移。因此,本文在统一模型与数据集设定、控制扰动不可感知性相近的前提下,选取当前已被广泛验证有效的多种代表性对抗攻击方法开展比较实验。在每个类别中均选取100张原始图像作为基准,系统评估并比较各方法在对抗样本生成效率方面的表现。表2、表3和表4为在扰动幅值严格受限、感知差异近似可控的同等条件下,本文方法与经典的快速梯度符号方法(Fast Gradient Sign Method, FGSM)^[23]、关注错误分类类别与原始类别距离的 NetSat^[24]方法在3个数据集上的 ALD_p 、ASS 和对抗效率指标差异。其中, ALD_p 值越小越好,而 ASS 值越大越好。可以看出,在保持不可感知性指标相近的情况下,本文方法的有效性相较于其他方法略有下降,这是由于本文方法的目标为仅对少量像素扰动并确保 DNN 输出严重错误,因此对抗有效性下降10%左右通常认为在合理范围内。

Table 2 Comparison of adversarial efficiency under similar imperceptibility of MNIST test cases

对抗算法	ALD_p	ASS	错误率/%
FGHD-adv	52.98	0.253	85.2
NetSat	53.13	0.082	27.8
FGSM	53.78	0.087	91.0

Table 3 Comparison of adversarial efficiency under similar imperceptibility of CIFAR-10 test cases

对抗算法	ALD_p	ASS	错误率/%
FGHD-adv	24.93	0.181	75.6
NetSat	25.44	0.152	79.2
FGSM	25.08	0.135	91.6

3.4.2 对抗能力生成实验

FGHD-adv算法中的参数 λ 可调节生成用例触发错误的分散程度。实验分别选取 λ 为2和0.1来生成测试用

Table 4 Comparison of adversarial efficiency under similar imperceptibility of STL-10 dataset

对抗算法	ALD _p	ASS	错误率/%
FGHD-adv	34.92	0.254	73.6
NetSat	35.25	0.236	76.8
FGSM	35.72	0.232	88.6

例,运行结果见图 2。可以看出,每张图中将不同模型得到的 SCSM 矩阵类别放置由在 x 轴和 y 轴组成的底面, z 轴表示测试用例对应输出某一标签的数量。 y 轴的第 0 列表示正确标签, y 值越大则类别差异越大,表示错误越严重。当

λ 较大时,加权高斯函数变得更为平缓,导致出现错误的结果类别更加远离原始类别。相反,当 λ 较小时,加权高斯函数变得更陡峭,表现为分类更靠近 y 轴 0 点的类别明显更多一些。

在不同的不可感知性下验证本文方法能否生成具有更大 ADM 指标的样本。取 λ 的加权值为 2.0,在不同数据集下,不同方法分别在一定范围的不可感知性指标内各生成 1 000 张测试用例,其 ADM 指标结果如图 3 所示。可以看出,FGHD-adv 算法生成的测试用例在不同大小的不可感知性指标下均有更好的 ADM 结果,说明其在 3 种数据集上均能很好地触发更严重的错误。

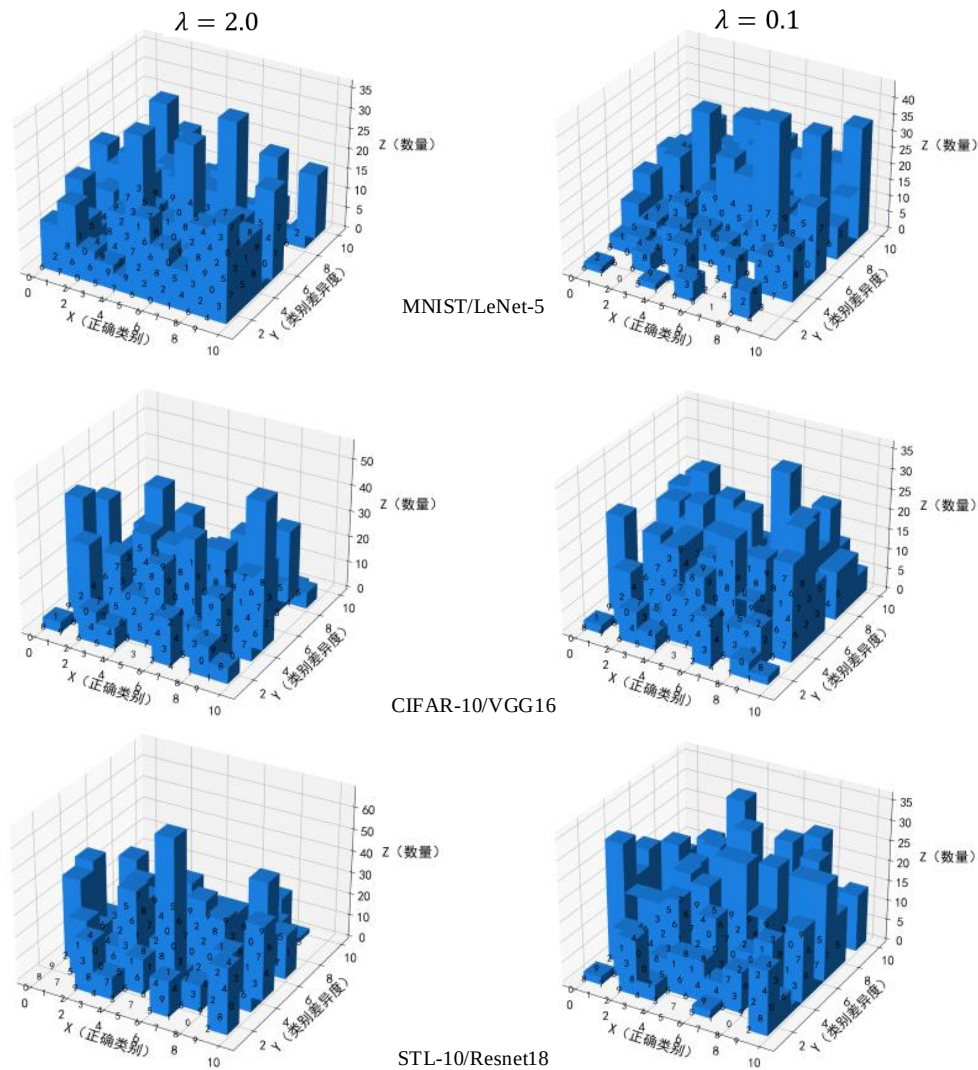


Fig. 2 Classification result category statistics chart
图 2 分类结果类别统计图

4 结语

本文针对高风险测试用例生成技术提出 3 方面改进:①提出错误严重程度指标 ADM 作为测试用例集风险度量;②利用核函数平滑 SCSM 矩阵,可按测试要求生成不同

严重等级的测试用例集;③利用关键像素提高测试用例的不可感知性。在多个数据集上的实验结果表明,在不可感知性相近的情况下,与其他方法相比,本文方法生成的测试用例集可触发更严重的错误。不足之处在于关键像素选取的计算量较大,未来将引入特征显著度作为关键像素的选取依据。

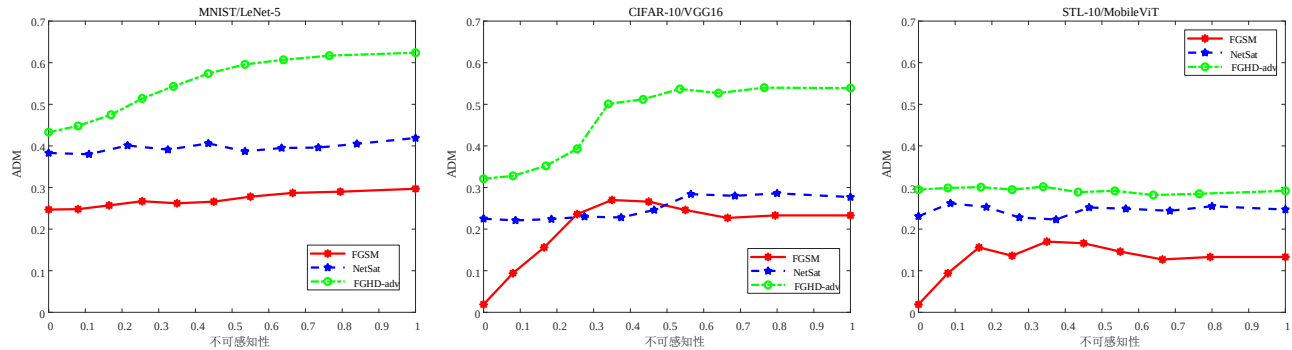


Fig. 3 Variation of ADM with imperceptibility

图3 ADM随不可感知性的变化

参考文献:

- [1] LU C C, HUANG L C, LIU J W. Research on diabetes prediction based on IG-DNN hybrid decision algorithm [J]. Software Guide, 2019, 18(8): 21-25.
卢春城, 黄理灿, 刘靖雯. 基于IG-DNN混合决策算法的糖尿病预测研究[J]. 软件导刊, 2019, 18(8): 21-25.
- [2] HUANG H B, WAN L, CHU K. Research on intrusion detection method based on Transformer [J]. Software Guide, 2022, 21(9): 121-128.
黄海彬, 万良, 褚望. 基于Transformer的入侵检测方法研究[J]. 软件导刊, 2022, 21(9): 121-128.
- [3] JAMONNAK S, ZHAO Y, HUANG X, AMIRUZZAMAN M, et al. Geo-context aware study of vision-based autonomous driving models and spatial video data [J]. IEEE Transactions on Visualization and Computer Graphics, 2022, 28(1): 1019-1029.
- [4] Tesla Motors Club. Search results for "phantom braking" on Tesla Motors Club forum [EB/OL]. <https://teslamotorsclub.com/tmc/search/13796300/?q=phantom+braking&o=relevance>.
- [5] The Dawn Project. Keep our children safe from Tesla self-driving [EB/OL]. <https://dawnproject.com/keep-our-children-safe-from-tesla-self-driving/>.
- [6] JEANNERET G, PÉREZ J C, ARBELÁEZ P, et al. A hierarchical assessment of adversarial severity [C]//IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), 2021: 61-70.
- [7] FILUS K, DOMAŃSKA J. Similarity-driven adversarial testing of neural networks [J]. Knowledge-Based Systems, 2024, 305: 112621.
- [8] LIU H, LI K X, CHEN Y X. Survey on trustworthiness measurement for artificial intelligence systems [J]. Journal of Software, 2023(8): 3774-3792.
刘晗, 李凯旋, 陈依香. 人工智能系统可信性度量评估研究综述[J]. 软件学报, 2023(8): 3774-3792.
- [9] AKRAM S, BAZAI S U, GHAFOR M I, et al. Systematic literature review: evaluating effects of adversarial attacks and attack generation methods [C]//International Conference on Energy, Power, Environment, Control, and Computing (ICEPECC), 2023: 1-6.
- [10] CHEN X L, ZAN D G, WU B C, et al. Adversarial example generation algorithm for vertical federated learning [J]. Journal on Communications, 2023, 44(8): 1-13.
陈晓霖, 曾道广, 吴炳潮, 等. 面向纵向联邦学习的对抗样本生成算法[J]. 通信学报, 2023, 44(8): 1-13.
- [11] JI S H, HU L M, ZHANG P C, et al. Adversarial example generation method based on sparse perturbation [J]. Journal of Software, 2023, 34(9): 4003-4017.
吉顺慧, 胡黎明, 张鹏程, 等. 基于稀疏扰动的对抗样本生成方法[J]. 软件学报, 2023, 34(9): 4003-4017.
- [12] SUN J Z, CHEN L, LEI L, et al. An adversarial robustness evaluation platform for deep learning models on Image classification [J]. Electronics, 2023, 12(17): 3665.
- [13] OGONNA M, ADENIRAN A, ADEYEMO A. Relationship between uncertainty in DNNs and adversarial attacks [DB/OL]. <https://doi.org/10.48550/arXiv.2409.13232>.
- [14] DENG L. The MNIST database of handwritten digit images for machine learning research [J]. IEEE Signal Processing Magazine, 2012, 29(6): 141-142.
- [15] LECUN Y, BOTTOU L, BENGIO Y, et al. Gradient-based learning applied to document recognition [J]. Proceedings of the IEEE, 1998, 86(11): 2278-2324.
- [16] University of Toronto. CIFAR-10 and CIFAR-100 datasets [EB/OL]. <https://www.cs.toronto.edu/~kriz/cifar.html>.
- [17] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition [DB/OL]. <https://doi.org/10.48550/arXiv.1409.1556>.
- [18] Stanford University. STL-10 dataset [EB/OL]. <https://cs.stanford.edu/~acoates/stl10/>.
- [19] MEHTA S, RASTEGARI M. MobileViT: light-weight, general-purpose, and mobile-friendly vision transformer [DB/OL]. <https://doi.org/10.48550/arXiv.2110.02178>.
- [20] YAN H, YANG F Y, ZHONG Y H, et al. Survey on test input selection and metric standards for deep neural networks [J]. Computer Engineering and Applications, 2024(6): 27-42.
严荏, 杨丰玉, 钟依慧, 等. 深度神经网络的测试输入选择与度量标准研究综述[J]. 计算机工程与应用, 2024(6): 27-42.
- [21] WANG Z, BOVIK A C, SHEIKH H R, et al. Image quality assessment: from error visibility to structural similarity [J]. IEEE Transactions on Image Processing, 2004, 13(4): 600-612.
- [22] LING X, JI S, ZOU J, et al. Deepsec: a uniform platform for security analysis of deep learning model [C]//Proceedings of the IEEE Symposium on Security and Privacy, 2019: 673-690.
- [23] GOODFELLOW I J, SHLENS J, ZEGEDYS C. Explaining and harnessing adversarial examples [DB/OL]. <https://doi.org/10.48550/arXiv.1412.6572>.
- [24] FILUS K, DOMAŃSKA J. Netsat: network saturation adversarial attack [C]//2023 IEEE International Conference on Big Data, 2023: 5038-5047.

(责任编辑:尹晨茹)