计算机科学
COMPUTER SCIENCE

基于局部梯度平滑的解释鲁棒性对抗训练方法

陈自刚, 潘鼎, 冷涛, 朱海华, 陈龙, 周由胜

引用本文

陈自刚, 潘鼎, 冷涛, 朱海华, 陈龙, 周由胜. 基于局部梯度平滑的解释鲁棒性对抗训练方法[J]. 计算机科学, 2025, 52(2): 374-379.

CHEN Zigang, PAN Ding, LENG Tao, ZHU Haihua, CHEN Long, ZHOU Yousheng. Explanation Robustness Adversarial Training Method Based on Local Gradient Smoothing [J]. Computer Science, 2025, 52(2): 374-379.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.240400210

基于局部梯度平滑的解释鲁棒性对抗训练方法

陈自刚[1,2,3]　潘　鼎[1]　冷　涛[2]　朱海华[1]　陈　龙[1]　周由胜[1]
1 重庆邮电大学网络空间安全监测与治理重庆市重点实验室　重庆 400065
2 四川警察学院智能警务四川省重点实验室　四川 泸州 646000
3 重庆邮电大学网络空间大数据智能安全教育部重点实验室 重庆 400065
(chenzg@cqupt.edu.cn)

摘　要　深度学习可解释性在发展的同时,也面临着安全性方面的巨大挑战。模型对输入数据的解释结果存在被恶意操纵攻击的风险,此攻击严重限制了可解释性技术的应用场景并阻碍了人类对模型的探索与认知。针对此问题,提出一种使用模型梯度作为相似性约束的解释鲁棒性对抗训练方法。首先,沿解释方向采样生成对抗训练数据;其次,结合训练过程中样本的梯度信息来计算采样数据解释之间的多种相似性指标,用以对模型正则化,平滑模型的曲率;最后,为验证所提出的解释鲁棒性对抗训练方法的有效性,在多个数据集和解释方法上进行验证,实验结果表明,所提方法在防御对抗解释样本上具有显著效果。

关键词:深度学习;可解释性;对抗攻击;对抗训练;对抗样本
中图分类号　TP309

Explanation Robustness Adversarial Training Method Based on Local Gradient Smoothing

CHEN Zigang[1,2,3],PAN Ding[1],LENG Tao[2],ZHU Haihua[1],CHEN Long[1] and ZHOU Yousheng[1]
1 Chongqing Key Laboratory of Cyberspace Security Monitoring and Governance, Chongqing University of Posts and Telecommunications, Chongqing 400065, China
2 Intelligent Policing Key Laboratory of Sichuan Province, Sichuan Police College, Luzhou, Sichuan 646000, China
3 Key Laboratory of Cyberspace Big Data Intelligent Security, Ministry of Education, Chongqing University of Posts and Telecommunications, Chongqing 400065, China

Abstract　While the interpretability of deep learning is developing, its security is also facing significant challenges. There is a risk that the interpretation results of the model on input data may be maliciously manipulated and attacked, which seriously affects the application scenarios of interpretability technology and hinders human exploration and cognition of the model. To address this issue, an interpretable robust adversarial training method using model gradients as similarity constraints is proposed. Firstly, adversarial training data is generated by sampling along the interpretation direction. Secondly, multiple similarity metrics between the interpretations of the sampled data are calculated by combining the gradient information of the samples during the training process, which is used to regularize the model and smooth its curvature. Finally, to verify the effectiveness of the proposed interpretable robust adversarial training method, it is validated on multiple datasets and interpretation methods. The experimental results show that the proposed method has a significant effect on defending against adversarial interpretation samples.

Keywords　Deep learning, Interpretability, Adversarial attack, Adversarial training, Adversarial samples

1 引言

机器学习技术在计算机视觉、自然语言处理、语音识别等核心领域取得了显著的进展。多种机器学习模型被广泛应用于人脸识别[1]、自动驾驶[2]、恶意软件检测[3]以及医学影像分析[4]等关键任务中,并在这些任务中展现出与人类相当甚至超越人类的表现水平。这些应用不仅提升了机器学习技术对数据的理解和处理能力,同时也为社会进步带来了巨大的便利。在计算机视觉领域,深度神经网络展现出卓越的性能,成功解决了图像分类[5]、目标检测[6]、语义分割[7]等复杂任务。自 2012 年 ImageNet 大规模视觉识别挑战赛[8]举办以来,众多优秀的图像分类模型如 VGGNet[9],ResNet[10] 和 DenseNet[11]等相继涌现,推动了深度神经网络在图像识别领域的效能持续提升。此外,深度神经网络的应用范围也逐渐拓展至工业、农业以及网络安全等多个领域,展现出强大的适应性和潜力。

到稿日期:2024-04-30　返修日期:2024-08-26
基金项目:国家自然科学基金(62272076);智能警务四川省重点实验室开放基金重点项目(ZNJW2022KFZD002)
This work was supported by the National Natural Science Foundation of China(62272076) and Opening Project of Intelligent Policing Key Laboratory of Sichuan Province(ZNJW2022KFZD002).
通信作者:冷涛(lengtao@iie.ac.cn)

然而，随着机器学习技术的不断发展，其面临的挑战和问题也日益凸显。特别是模型复杂度的提升和大规模数据处理需求的增加，使得机器学习模型的“黑盒”属性愈发明显，进而引发了对其可信度和安全性的担忧。这种“黑盒”属性指的是模型内部运作机制和决策逻辑的不透明性，导致人们难以解释和验证模型的输出结果，从而降低了对模型的信任度。各国组织对深度学习可解释性的重视程度不断提高。我国《人工智能白皮书(2022)》[12]强调了模型可解释性的重要性，并提出了加强神经网络可解释性的需求。同样，欧盟《通用数据保护条例》也将透明度作为数据处理的关键原则之一，要求人工智能决策需具备可质疑性[13]。

随着可解释性研究的快速发展，研究者开始考虑可解释性的安全问题。Ghorbani 等[14]讨论了神经网络特征重要性解释方法的脆弱性，并提出了 Top-k、目标攻击和质心攻击 3 种攻击策略，通过简单梯度[15]、集成梯度[16]和 DeepLIFT[17]等解释方法验证了攻击的有效性。Dombrowski 等[18]则从理论上探讨了神经网络模型的几何特性与决策解释之间的联系，认为解释方法的脆弱性可能与模型的曲率特性相关。他们设计了一种基于优化的攻击策略来测试基于梯度的解释方法，并提出了通过平滑模型梯度曲率来增强解释鲁棒性的防御方案。Zhang 等[19]指出了模型决策与解释之间并非完全耦合的问题，并介绍了一种名为 ADV2 的新型攻击方法。该方法能够同时欺骗深度神经网络及其解释方法，使对抗性输入误导分类结果并欺骗解释器生成特定的目标归因图，实现对 DNNs 及其解释器的全面攻击。与此同时，Heo 等[20]提出了一种观点，即在攻击者能够修改模型权重的情况下，其攻击手段将对所有解释方法有效。然而，这一假设可能过于强大，在实际对抗环境中实现难度过高。本文更关注一个更实际的场景：攻击者只能修改输入数据而无法改动网络结构。

为解决可解释性容易受到恶意操纵这一问题，提出基于局部梯度平滑的解释鲁棒性对抗训练方法。该方法通过沿解释方向采样生成对抗训练数据，结合训练过程中样本的梯度信息来计算采样数据解释之间的相似性，用以对模型正则化，平滑模型的曲率，最终实现模型对解释操纵攻击的防御能力提升，增强模型鲁棒性。

2 解释鲁棒性对抗训练方案

2.1 对抗训练及相关理论定义

深度学习对抗训练是一种有效的提高模型鲁棒性的方法。该方法通过引入对抗样本，即在原始输入数据上添加微小扰动以误导模型，来增强模型对噪声和攻击的抵抗能力。在对抗训练中，模型不仅学习对正常样本的正确分类，还学习对对抗样本的鲁棒性，从而提高了其在复杂环境下的泛化性能。面向分类准确性的对抗训练思想如式(1)所示：

$$\min_f \alpha_1 \cdot \sum_{x\in\mathcal{D}} L(f(x),y)+\alpha_2 \cdot \sum_{x'\sim\mathcal{N}(x,\epsilon)} L(f(x'),y) \tag{1}$$

其中，x 和 y 为输入数据和真实标签，x' 为输入数据在扰动范围$(-\epsilon,\epsilon)$内的对抗样本，$L(\cdot,\cdot)$为真实标签与预测分数之间的距离，α_1 和 α_2 表示控制良性数据损失值和对抗攻击数据损失值之间的强度参数。

面向解释的对抗训练的目的却有所不同，面向解释的对抗训练不仅需要保证模型原本的分类性能，还需要保证模型对预测结果的解释足够鲁棒。一种自然的训练思想为在式(1)的基础上增加关于解释的损失项，如式(2)所示：

$$\min_f \alpha_1 \cdot \sum_{x\in\mathcal{D}} L(f(x),y)+\sum_{x'\sim\mathcal{N}(x,\epsilon)} [\alpha_2 \cdot L(f(x'),y)+\alpha_3 \cdot d(\phi(f_y,x'),\phi(f_y,x))] \tag{2}$$

其中，$\phi(f_y,x)$表示模型 f 在输入数据 x 上对预测结果 y 的解释结果。此对抗训练理论上可以对解释操纵起到防御效果，但随之而来的是训练成本的提升，对抗训练需要计算额外的解释。针对如 Grad-CAM++[21]这类计算成本较大的解释时，对抗训练的效率会进一步下降，所以对抗训练效率是对抗训练中非常重要的评判因素。

解释鲁棒性指模型在各种对抗或非对抗环境中具备稳定解释能力，无论输入数据是否为恶意对抗解释样本，模型都能如实输出真实的解释结果。假设目标模型为 $f:R^D\rightarrow R^C$，其中 $x\in R^D$，C 是生成解释所指定的类别(通常为预测分数最高的标签类别)。模型 f 对于输入的样本数据 x 的解释记作 $\phi(f_c,x)$。

为了便于理论分析和降低获得解释的计算成本，以下的理论分析选择经典的关于梯度的解释方法 Grad 作为对抗训练的默认解释方法。Grad 的计算方式如式(3)所示：

$$\phi(f_c,x)=\nabla_x f_c(x) \tag{3}$$

设一个添加微小对抗扰动噪声的样本为 $x'=x+\Delta x$，则解释的变化可以记作 $\Delta\phi$，其可以理解为模型对于解释的稳定性，$\Delta\phi$ 越小代表模型的解释越稳定。由 Grad 解释方法可以推理 $\Delta\phi$ 为：

$$\Delta\phi=\phi(f,x+\Delta x)-\phi(f,x)=\boldsymbol{H}\Delta x+O(|\Delta x|^2) \tag{4}$$

其中，$\boldsymbol{H}$ 表示海森矩阵，即 $\boldsymbol{H}_{i,j}=\dfrac{\partial f}{\partial x_i \partial x_j}$，这意味着如果函数 f 是线性的，则 f 的二阶梯度为零，$\Delta\phi$ 也接近于零，说明其具有非常高的解释稳定性，对任何解释操纵都是鲁棒的。然而，深度神经网络引入非线性的激活函数，虽然可以更好地拟合复杂任务的函数，但也使深度神经网络不能达到这种理想状态。故对深度神经网络的解释鲁棒性定义如式(5)所示：

$$\min_{\gamma>0} \max_{\Delta x} |\phi(f,x+\Delta x)-\gamma\phi(f,x)_2 \tag{5}$$

因为解释结果是根据特征之间的相对重要性生成的，所以在解释鲁棒性定义中设置 γ 系数不影响解释之间的相对关系。假设函数 f 对于样本 x 具有解释鲁棒性，则可以描述为 $f(x)=\sigma(\phi^{\mathrm{T}}x)$，其中 σ 是一个非线性函数，函数的权重由解释向量组成。结合 Grad 解释方法的定义并对 $f(x)$求梯度，可以得到 $\phi(f,x)=\sigma'(\phi^{\mathrm{T}}x)\cdot\phi$，而由于$\sigma'(\phi^{\mathrm{T}}x)$是一个标量，因此可以理解为 Δx 只是对具有解释鲁棒性的函数 f 的解释结果做了一次缩放操作，而缩放操作不会影响特征之间的相对稳定性。故将解释鲁棒性定义中的系数 γ 设置为$\sigma'(\phi^{\mathrm{T}}x)$。

2.2 数据采样方法

对于计算复杂度较高的解释方法，其对抗样本获取成本过高，对抗训练效率低下，且由于 ReLU 激活函数的特性，其二阶梯度难以计算，这严重影响了解释鲁棒性的训练效果。对抗训练可被视为一种特殊的使用对抗样本进行的数据增强技术，如何制作对抗训练数据集是对抗训练的核心。我们注意到，模型的解释结果可以一定程度上代表模型的注意力关注点，所以一种直观的推断是：样本在沿解释方向移动时不会

对预测的准确度和解释结果产生影响；相反，如果样本移动的方向偏离了解释方向，那么预测的准确性和解释准确性就会降低。本文假设垂直于解释方向上的增强数据对解释的稳定性影响最大。受此启发，沿解释方向移动的样本也可作为一种数据增强手段用于对抗训练。我们可以通过缩小沿解释样本与沿垂直解释样本之间的预测分数差距来提高目标模型的解释鲁棒性。沿解释方向的数据采样方法如式(6)所示：

$$x^i = x + \delta_1 \phi(f,x) / \| \phi(f,x) \|_2 \quad \text{s.t. } -\Delta_1 \leqslant \delta_1 \leqslant \Delta_1 \tag{6}$$

其中，x 表示原输入数据；δ_1 是数据沿解释方向移动的偏移量；$\phi(f,x)$ 表示模型对输入数据的解释结果；$\| \cdot \|_2$ 表示解释的欧几里得距离，用以对解释归一化。这里 ϕ 选择使用 Grad 方法来简单高效地计算解释。此外，垂直于解释方向的采样方法如式(7)所示：

$$x^p = x^i + \delta_2 \phi_\perp(f,x) / \| \phi_\perp(f,x) \|_2 \quad \text{s.t. } -\Delta_2 \leqslant \delta_2 \leqslant \Delta_2 \tag{7}$$

其中，$\phi_\perp$ 表示垂直方向的解释，其计算方式为 $\phi_\perp = u - \phi \cdot \langle u, \phi \rangle / |\phi|_2^2$，$u$ 是一个均匀分布。垂直于解释方向的样本可以有效改变原数据中梯度的方向和大小，使基于梯度的解释方法产生偏移。

2.3 损失函数设计

本节将讨论所提对抗训练的损失函数的设计构思。损失函数的设计主要分为 4 个部分，为保证模型原本的预测性能，第一部分为模型训练前的原损失函数，通常设置为交叉熵损失函数，表示为：

$$L_1 = -\sum_{i=1}^{n} y_i \log f(x_i) \tag{8}$$

其中，n 为训练数据的样本量。第二部分为数据采样部分的正则项，主要目的是计算沿解释方向预测分数与沿垂直解释方向预测分数的接近程度。为了使两个采样数据的预测分数一致，同样损失设置为交叉熵损失，记为：

$$L_2 = -\sum_{x^i \sim J(x) x^p \sim \mathcal{P}(x^i)} f(x^i) \log f(x^p) \tag{9}$$

为了进一步增强 x^i 和 x^p 在梯度大小和方向上的一致性，进一步平滑模型的决策边界，对抗训练采用数据点之间梯度的相似性作为正则项。

在相似性指标的选择上，本文采用 L_2 相似性和余弦相似性作为损失正则项。

L_2 距离，也被称为欧几里得距离，是度量空间中两个向量之间直线距离的一种方法。它通过计算两个向量在对应维度上差值的平方和，并取其平方根来得到它们之间的几何距离。当 L_2 距离较小时，表明向量间的相似度较高；相反，较大的 L_2 距离则意味着向量间存在较大的差异。L_2 正则项计算 x^i 和 x^p 两个采样数据之间的梯度的欧几里得距离，其计算过程如式(10)所示：

$$L_{l_2} = \| \nabla f(x^i) - \nabla f(x^p) \|_2 \tag{10}$$

其中，x^i 表示输入数据沿解释方向上的采样数据，x^p 表示在垂直解释方向上的采样数据。

余弦相似性通过计算两个向量间夹角的余弦值来衡量它们之间的相似程度。当两个向量的夹角为 0 时，即它们方向完全相同时，余弦值为 1，表明这两个向量完全相似。而夹角为 180°，即两个向量方向完全相反时，余弦值为 −1，表明它们完全不相似。因此，余弦相似性的取值范围在 −1～1 之间，值越大表示两个向量越相似，具体计算过程如式(11)所示：

$$cossim(A,B) = \frac{A \cdot B}{\| A \| \| B \|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \times \sqrt{\sum_{i=1}^{n} (B_i)^2}} \tag{11}$$

为了更好地计算采样数据之间的梯度相似性，本文将余弦相似度归一化，使越相似的梯度损失越小。余弦正则项的计算过程如式(12)所示：

$$L_{\cos} = \frac{1}{2}(1 - cossim(\nabla f(x^i), \nabla f(x^p))) \tag{12}$$

综上，对抗训练的总损失函数如式(13)所示：

$$\begin{aligned} Loss &= L_1 + L_2 + L_{l_2} + L_{\cos} \\ &= -\sum_{a=1}^{n} y_a \log f(x_a) + \alpha_1 \cdot \sum_{x^i \sim J(x) x^p \sim \mathcal{P}(x^i)} f(x^i) \log f(x^p) + \alpha_2 \cdot \| \nabla f(x^i) - \nabla f(x^p) \|_2 + \alpha_3 \cdot \frac{1}{2}(1 - cossim(\nabla f(x^i), \nabla f(x^p))) \end{aligned} \tag{13}$$

其中，α_1，α_2，α_3 均为控制正则项强度的超参数。基于以上理论，解释鲁棒性对抗训练的具体算法流程如图 1 所示。

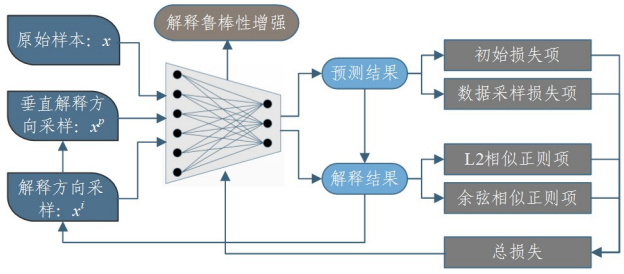


图 1 解释鲁棒性对抗训练框架

Fig. 1 Explaining robustness adversarial training framework

3 实验与分析

3.1 数据及实验设置

实验在云服务器上进行，CPU 型号为 AMD EPYC 7T83，GPU 型号为 NVIDIA RTX 4090，深度学习框架为 PyTorch 1.11.0，开发语言为 Python3.8(Ubuntu20.04)，驱动环境为 Cuda 11.3，第三方库包括 numpy pandas matplotlib 等。

实验选用 MNIST，Fashion-MNIST 和 CIFAR-10 这 3 个经典的数据集来进行验证。MNIST 是一个手写数字识别数据集，包含了 60000 个训练样本以及 10000 个测试样本，每个样本均为 28×28 像素的灰度图像，展示了从 0－9 的手写数字。Fashion-MNIST 数据集作为 MNIST 的扩展，涵盖了 10 个类别的时尚单品图像，例如 T 恤、裤子、鞋子等，其同样包含 60000 个训练样本和 10000 个测试样本，每个样本也均为 28×28 像素的灰度图像。CIFAR-10 数据集是一个包含 10 个类别的彩色图像数据集，其中涉及飞机、汽车、鸟类、猫等自然和人造对象。CIFAR-10 数据集共包含 60000 个 32×32 像素的彩色图像，其中 50000 个图像被指定为训练样本，10000 个图像作为测试样本。相较于 MNIST 和 Fashion-MNIST，CIFAR-10 数据集的图像质量、大小和颜色都更为复杂。实验选用这 3 个各具特色的数据集，旨在全面评估所提出的

对抗训练方法的性能，从而验证其在不同图像分类任务中的有效性和优越性。

3.2 评价指标

评价指标方面，本节除了采用模型准确率外，还采用随机扰动相似性（Random Perturbation Similarity，RPS）[22]和插入游戏[23]来评估模型生成解释的鲁棒性。

RPS 主要用于衡量原始数据 x 和随机扰动数据 $x+\delta_x$ 之间解释的相似性，其主要思想是在原始数据 x 和随机扰动数据 $x+\delta_x$ 上添加随机扰动，并观察 x 和 $x+\delta_x$ 被扰动后生成解释之间的相似性。RPS 越高表示所添加扰动 δ_x 对解释的影响越低，模型的解释鲁棒性越高。

对于给定的数据集 $\mathcal{D}$、解释算法 ϕ、相似性评价指标 S 和随机扰动范围 ϵ，RPS 的计算过程如式(14)所示：

$$RPS=\frac{1}{N}\sum_{x,y\in\mathcal{D}}S(\phi(f_y,x),\phi(f_y,x+\delta_x)) \tag{14}$$

其中，N 表示数据集 $\mathcal{D}$ 中数据元组的数量；δ_x 表示在 $[-\epsilon,\epsilon]$ 区间的随机扰动噪声。实验中选择使用余弦相似度、皮尔逊相关系数（Pearson Correlation Coefficient，PCC）和结构相似性指数（Structural Similarity，SSIM）作为相似性评价指标 S。余弦相似度计算过程已在 3.1 节介绍，此处不再赘述。PCC 广泛用于度量两个变量之间的相关程度，其值介于 $-1\sim1$ 之间，绝对值越大表明相关性越强。两个变量之间的皮尔逊相关系数定义为两个变量之间的协方差和标准差的商，其具体计算式如式(15)所示：

$$PCC(X,Y)=\frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i-\bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i-\bar{Y})^2}} \tag{15}$$

相对于其他评价指标，SSIM 旨在更准确地评估两幅图像的视觉相似性。SSIM 基于图像的亮度、对比度和结构信息来综合评价图像间的相似性，它通过计算两幅图像在局部窗口内的亮度相似性、对比度相似性和结构相似性的加权乘积来得到最终的 SSIM 指数。这 3 个因素分别反映了图像在亮度层次、对比度差异和结构信息上的相似性，其计算式如式(16)所示：

$$SSIM(x,y)=\frac{(2\mu_x\mu_y+c_1)(2\sigma_{xy}+c_2)}{(\mu_x^2+\mu_y^2+c_1)(\sigma_x^2+\sigma_y^2+c_2)} \tag{16}$$

其中，μ_x 为 x 的均值，μ_y 为 y 的均值，σ_x 和 σ_y 为 x 和 y 的方差，σ_{xy} 为 x 和 y 的协方差。

插入游戏是一种衡量归因解释质量的经典评估方法，其主要思想是按解释结果来重构输入像素并观察输出分数的变化，按归因解释的得分顺序插入来观察真实标签 y 的分类概率的变化。在插入游戏中，如果解释结果偏离真实解释注意力区域，那么插入游戏分数将会降低。在插入游戏中设置参数 γ 表示按解释得分 $\phi(f_y,x)$ 从零值图像 x_0 重建输入元素的比率。重建图像的过程中将设置一个掩码 $m_\gamma^J\in\{0,1\}^d$，用于遮盖解释得分低于某个阈值的区域，每个元素均满足 $(m_\gamma^J)_i=1[\phi(f_y,x)_i>t_\gamma]$，其中 t_γ 的阈值计算满足 $\sum_i(\boldsymbol{m}_\gamma^J)_i/d=\gamma$。因此，重构后的输入记作 $x_\gamma^J=x\odot\boldsymbol{m}_\gamma^J+x_0\odot(1-\boldsymbol{m}_\gamma^J)$，插入游戏计算的平均预测分数如式(17)所示：

$$Insertion_g(J,\mathcal{D},\gamma)=\frac{1}{N}\sum_{(x,y)\in\mathcal{D}}p_y(x_\gamma^J) \tag{17}$$

3.3 实验结果分析

首先评估所提对抗训练方法的训练效果。实验在 MNIST，Fashion-MNIST 和 CIFAR-10 这 3 个数据集上进行。采用 Grad，Input×Grad，GuidedBP[24] 和 LRP[25] 这 4 种解释算法进行对抗训练。这些解释算法都会生成像素级显著图的解释结果，像素的分布和数值表示与神经网络预测分数的相关程度。实验采用 3 个卷积＋池化层和 2 个全连接层的神经网络进行验证。由于我们的验证目标不是为了在这些数据集上达到最先进的预测准确率，而是提升模型的解释鲁棒性，因此在所有的对抗训练中统一将模型的激活函数按文献[18]设置为 Softplus 函数，其中 β 设置为 3，这种设置可以使模型的几何形状更加平滑，以增强解释鲁棒性。在模型进行对抗训练时，采用交叉熵损失作为基线损失函数，对抗训练后的模型允许有 1%的分类准确度损失。

在对抗训练完成后，使用 3.1 节介绍的随机扰动相似性和插入游戏来评估模型对解释操纵的防御能力。

经过对随机扰动相似性的细致评估，可以得出如图 2 所示的综合结果，4 个子图分别给出了 Grad，Input×Grad，GuidedBP 和 LRP 这 4 种不同方法的评估情况。每个子图展示了在 MNIST，Fashion-MNIST 和 CIFAR-10 这 3 个广泛使用的数据集上的实验结果。

横轴上，Raw 是原始模型，该模型仅使用交叉熵作为损失函数进行训练，没有引入任何额外的正则化或对抗训练策略。ATEX，Hessian，LGA 和 RobustAGA 是为了进行对照实验而选择的解释鲁棒性对抗训练方法，这些方法在之前的研究中已被证明在提升模型对噪声和扰动的鲁棒性方面具有一定效果。

从图 2 中可以清晰地看到，本文所提出的方法（Ours）在各数据集和指标上都表现出了优异的性能。具体来说，当对图像进行随机扰动时，该方法所生成的解释与原图像的解释保持高度的一致性，这表明该方法在面对普通噪声时具有很强的抵抗力。

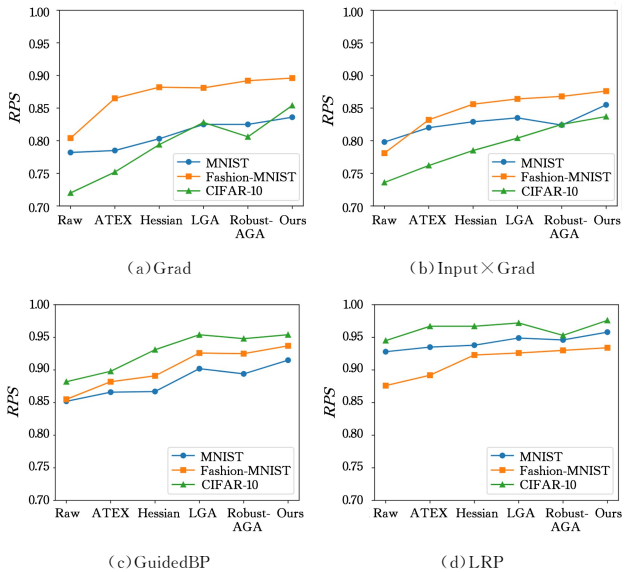


图 2 RPS 鲁棒性评估结果

Fig. 2 RPS robustness evaluation results

在插入游戏的评估环节，我们深入探究了模型在良性

样本和对抗解释样本中的性能差异。针对良性样本的评估结果如图 3 所示，我们呈现了 4 种解释方法 Grad，Input×Grad，GuidedBP 和 LRP 在多个数据集上的表现。从图中可以观察到，除了在 MNIST 数据集上 Input×Grad 的评估结果稍显逊色外，其余数据集在各解释方法下均展现出了令人满意的性能。这一结果表明，所提模型在良性样本上具有良好的泛化能力和解释性，特别是在处理复杂数据集时，模型能够准确地捕捉到数据的关键特征，并给出合理的解释。总体而言，良性样本的评估结果充分证明了所提模型在多数情况下都能够提供准确、可靠的解释，这为我们在对抗解释样本中的进一步探索奠定了坚实的基础。

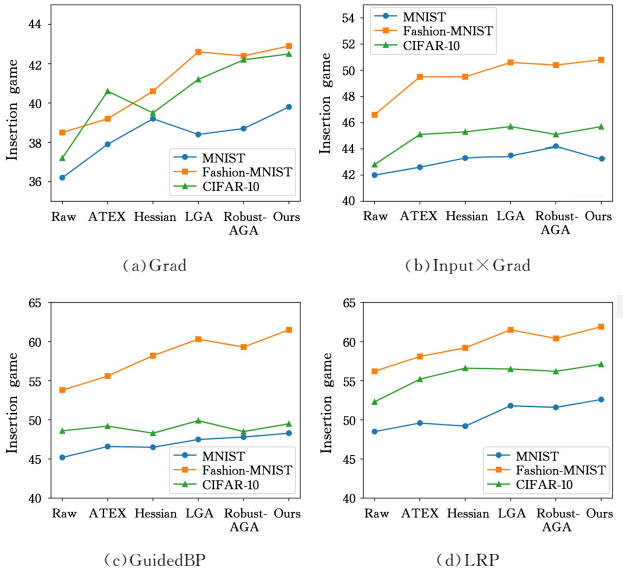


图 3 插入游戏评估结果(良性样本)

Fig. 3 Insert game evaluation results(benign samples)

插入游戏在使用对抗解释样本评估时的表现如图 4 所示。

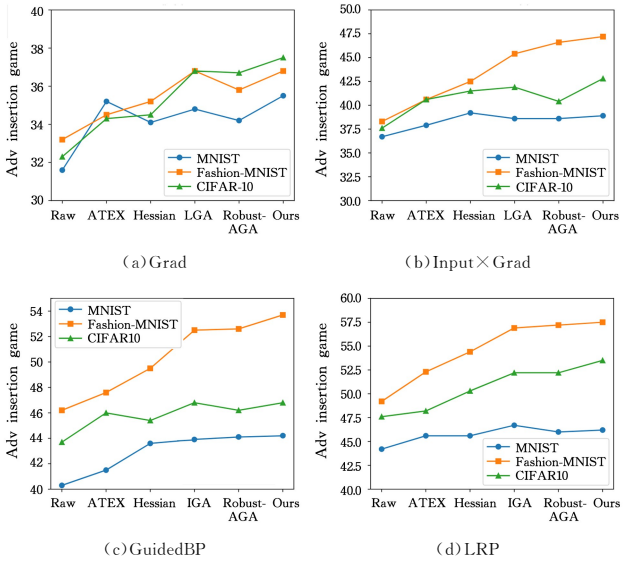


图 4 插入游戏评估结果(对抗解释样本)

Fig. 4 Insert game evaluation results(adversarial explanation samples)

4 种解释方法(Grad，Input×Grad，GuidedBP 和 LRP)在 MNIST 数据集上的评估结果相较于其他数据集均呈现一定程度的下滑。这表明在 MNIST 这一特定环境下，所提方法可能面临挑战。尽管如此，仔细观察数据可以发现，除了 Grad 子图中使用 ATEX 训练的模型以及 Input×Grad 子图中使用 Hessian 训练的模型在性能上略优于本文提出的对抗训练方法外，在其余所有实验环境和条件下，本文所提出的方法均展现出了显著的优势。这一发现有力地证明了在面对对抗解释操纵时，所提方法具备较强的防御能力。此外，我们还注意到，在多数实验场景中，本文提出的对抗训练方法不仅能够有效抵御对抗样本的干扰，还能为多种解释方法提供更强的鲁棒性。这意味着在实际应用中，所提方法有望帮助模型在面对复杂和潜在的攻击时保持更加稳定和可靠的性能。

对抗训练的效率是评估对抗训练方法实用性和可行性的核心指标之一。高效的训练过程不仅意味着模型能够更快速地获得所需的鲁棒性，还直接关联到实际应用中的时间成本和资源消耗。因此，在评估不同对抗训练方法时，我们也进行了训练效率的评估。

在实验中，我们以原始模型的训练时间作为基准，将其相对训练时间设定为 1，以便更直观地展示其他方法的训练时间与原始模型之间的比例关系。其他方法的训练时间均以原始模型训练时间的倍数来表示，这样的设置有助于清晰地对比不同方法的效率差异。

对比实验的结果如图 5 所示。可以明显看出，不同数据集和不同对抗训练方法之间的训练效率存在显著差异。具体而言，MNIST 数据集由于图案相对简单且规模较小，因此整体训练效率最高。相反，CIFAR-10 数据集则因为包含更复杂的图像内容和更大的数据量，因此整体训练效率最低。

在各对抗训练方法中，除了 ATEX 在整体效率上略优于本文提出的方法外，其他方法的整体效率均不及本文方法。这一结果表明，本文提出的对抗训练方法在实现模型鲁棒性的同时，也具有较高的训练效率。这一优势使得该方法在实际应用中更具吸引力和可行性，尤其是在对训练时间和资源消耗有严格要求的应用场景中。

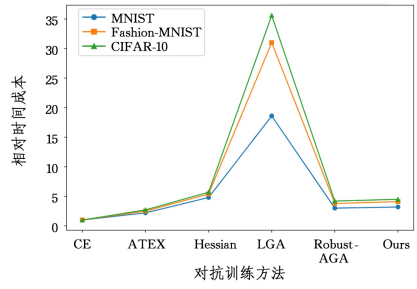


图 5 训练时间对比

Fig. 5 Training time comparison

结束语 本文提出了一种基于局部梯度平滑的解释鲁棒性对抗训练方法。该方法利用解释方法对输入数据进行多方向采样并将其作为数据增强方法，构造对抗训练数据集，并在对抗训练的损失函数中增加数据增强损失、余弦相似度和L_2距离相似性作为损失函数的正则项，来进一步约束模型的几何形状，使模型的局部梯度大小和方向都更加平滑。实验在 MNIST，Fashion-MNIST 和 CIFAR-10 数据集上进行，在

对抗训练中分别采用 Grad，Input × Grad，GuidedBP 和 LRP 解释方法对输入数据进行采样，并采用多种解释性对抗训练方法进行对比验证。实验结果表明，本文提出的基于局部梯度平滑的解释鲁棒性对抗训练方法可以有效提高模型的解释稳定性，并提高模型对于对抗解释操纵的解释鲁棒性。

参 考 文 献

[1] SCHROFF F，KALENICHENKO D，PHILBIN J. Facenet：A unified embedding for face recognition and clustering[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston，USA，2015：815-823.

[2] BOJARSKI M，DEL TESTA D，DWORAKOWSKI D，et al. End to end learning for self-driving cars[J]. arXiv：1604. 07316，2016.

[3] TOBIYAMA S，YAMAGUCHI Y，SHIMADA H，et al. Malware detection with deep neural network using process behavior [C]// 2016 IEEE 40th Annual Computer Software and Applications Conference(COMPSAC). Atlanta，USA，2016：2：577-582.

[4] WANG J，LI J Z，WANG Z T，et al. An Interpretable Prediction Model for Heart Disease Risk Based on Improved Whale Optimized LightGBM[J]. Journal of Beijing University of Posts and Telecommunications，2023，46(6)：39-45.

[5] KRIZHEVSKY A，SUTSKEVER I，HINTON G E. Imagenet classification with deep convolutional neural networks[J]. Communications of the ACM，2017，60(6)：84-90.

[6] REN S，HE K，GIRSHICK R，et al. Faster r-cnn：Towards real-time object detection with region proposal networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence，2017，39(6)：1137-1149.

[7] CHEN L C，PAPANDREOU G，KOKKINOS I，et al. Deeplab：Semantic image segmentation with deep convolutional nets，atrous convolution，and fully connected crfs[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence，2017，40(4)：834-848.

[8] RUSSAKOVSKY O，DENG J，SU H，et al. Imagenet large scale visual recognition challenge[J]. International Journal of Computer Vision，2015，115：211-252.

[9] SIMONYAN K，ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[J]. arXiv：1409. 1556，2014.

[10] HE K M，ZHANG X Y，REN S Q，et al. Deep residual learning for image recognition[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas，USA，2016：770-778.

[11] HUANG G，LIU Z，VAN DER MAATEN L，et al. Densely connected convolutional networks[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu，USA，2017：4700-4708.

[12] China Academy of Information and Communications Technology. Artificial Intelligence White Paper(2022) [R]. Beijing：China Academy of Information and Communications Technology，2022.

[13] REGULATION P. General data protection regulation [J]. Intouch，2018，25：1-5.

[14] GHORBANI A，ABID A，ZOU J. Interpretation of neural networks is fragile[C]// Proceedings of the AAAI Conference on Artificial Intelligence. Hawaii，USA，2019：3681-3688.

[15] SIMONYAN K，VEDALDI A，ZISSERMAN A. Deep inside convolutional networks：Visualising image classification models and saliency maps[J]. arXiv：1312. 6034，2013.

[16] SUNDARARAJAN M，TALY A，YAN Q. Axiomatic attribution for deep networks[C]// International Conference on Machine Learning. Sydney，Australia，2017：3319-3328.

[17] SHRIKUMAR A，GREENSIDE P，KUNDAJE A. Learning important features through propagating activation differences [C]// International Conference on Machine Learning. Sydney，Australia，2017：3145-3153.

[18] DOMBROWSKI A K，ALBER M，ANDERS C，et al. Explanations can be manipulated and geometry is to blame[C]// Proceedings of the 33rd International Conference on Neural Information Processing Systems. 2019：13589-13600.

[19] ZHANG X Y，WANG N F，SHEN H，et al. Interpretable deep learning under fire[C]// 29th {USENIX} Security Symposium ({USENIX} Security 20). Virtual Event，2020：1659-1676.

[20] HEO J，JOO S，MOON T. Fooling neural network interpretations via adversarial model manipulation[C]// Proceedings of the 33rd International Conference on Neural Information Processing Systems. 2019：2925-2936.

[21] CHATTOPADHAY A，SARKAR A，HOWLADER P，et al. Grad-cam＋＋：Generalized gradient-based visual explanations for deep convolutional networks[C]// 2018 IEEE Winter Conference on Applications of Computer Vision (WACV). Lake Tahoe，USA，2018：839-847.

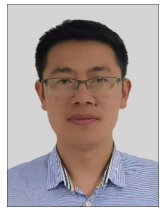
[22] DOMBROWSKI A K，ANDERS C J，MÜLLER K R，et al. Towards robust explanations for deep neural networks[J]. Pattern Recognition，2022，121：108194.

[23] PETSIUK V，DAS A，SAENKO K. Rise：Randomized input sampling for explanation of black-box models[J]. arXiv：1806. 07421，2018.

[24] SPRINGENBERG J T，DOSOVITSKIY A，BROX T，et al. Striving for simplicity：The all convolutional net[J]. arXiv：1412. 6806，2014.

[25] BACH S，BINDER A，MONTAVON G，et al. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation[J]. PloS One，2015，10(7)：e0130140.

CHEN Zigang，born in 1978，Ph.D，associate professor，is a member of CCF(No. F8469M). His main research interests include Internet of Things，Internet of Vehicles security and forensics，intelligent security and forensics，data security and privacy protection.



LENG Tao，born in 1986，Ph.D，associate professor. His main research interests include threat hunting，advanced threat detection，forensic analysis and graph neural networks.

（责任编辑：何杨）