基于决策边界光滑度的深度学习模型对抗鲁棒性评估指标

吴 涛,汪俊杰,曹新汶,王 练,先兴平,张睿康
(重庆邮电大学网络空间安全与信息法学院,重庆 400065)

摘 要: 深度学习模型的对抗鲁棒性对于可信人工智能发展至关重要. 研究领域广泛采用对抗攻击方法间接评价模型的对抗鲁棒性,然而此类方式依赖具体的对抗攻击方法和对抗扰动程度,无法反映模型的本质特征. 同时,仅有的少数直接进行模型对抗鲁棒性评价的评估指标要求对抗扰动的先验知识或者假设训练数据服从特定分布,适用性不强. 基于此,从模型自身特性出发,本文提出一种简单有效的、基于决策边界光滑度的对抗鲁棒性评估指标DBSE(Decision Boundary Shannon Entropy). 此方法利用对抗鲁棒性与决策边界光滑性之间的相关性,提出用于获取边界样本以近似刻画模型实际决策边界的"决策空间搜索策略". 然后,利用奇异值分解提取近似决策边界空间结构信息,并采用香农熵进行分布的均匀性量化,从而形成对抗鲁棒性评估指标DBSE. 实验结果表明,DBSE与代表性评估指标ASR(Attack Success Rate)、EBD(Empirical Boundary Distance)、ACTC(Average Confidence of True Class)、ACAC(Average Confidence of Adversarial Class)、MP(Minimal Perturbation)和ROBY相比,在独立性、有效性和时效性方面具有更好的表现,且不依赖对抗攻击方法,在时间开销方面比EBD减少了55%.

关键词: 鲁棒性评估;对抗鲁棒性;决策边界;对抗攻击;模型鲁棒性

基金项目: 国家自然科学基金(No.62376047,No.62106030);重庆市自然科学基金创新发展联合基金重点项目(No.CSTB2023NSCQ-LZX0003);重庆市教委科学技术研究计划重点项目(No.KJZD-K202300603);重庆市技术创新与应用发展面上项目(No.CSTB2022TIAD-GPX0014);中国高校产学研创新基金项目(No.2022BL105)

中图分类号: TP391 **文献标识码:** A **文章编号:** 0372-2112(2025)06-2090-14

电子学报URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20240932

Evaluation Metrics for Adversarial Robustness Based on the Smoothness of Decision Boundary in Deep Learning Models

WU Tao, WANG Jun-jie, CAO Xin-wen, WANG Lian, XIAN Xing-ping, ZHANG Rui-kang
(*School of Cyberspace Security and Information Law*, *Chongqing University of Posts and Telecommunications*, *Chongqing* 400065, *China*)

Abstract: The adversarial robustness of deep learning models is crucial for the development of trustworthy artificial intelligence. The research field widely adopts adversarial attack methods to indirectly evaluate the adversarial robustness of models. However, such methods rely on specific adversarial attack methods and levels of adversarial perturbations, failing to reflect the essential characteristics of models. Meanwhile, the few existing indicators that directly assess model adversarial robustness require prior knowledge of adversarial perturbations or assume that training data follows a specific distribution, limiting their applicability. In response to these challenges, starting from the intrinsic characteristics of models, this paper proposes a simple and effective adversarial robustness evaluation metric, DBSE. This method exploits the correlation between adversarial robustness and decision boundary smoothness, proposing a decision boundary sample sampling strategy to approximate and characterize the actual decision boundary of the models by obtaining samples about the decision boundary. Then, singular value decomposition is used to extract spatial structural information of the decision boundary, and Shannon entropy is employed to quantify the distribution of variations in various directions, thereby forming the adversarial robustness evaluation metric DBSE. Experimental results demonstrate that DBSE outperforms representative evaluation metrics such as ASR(Attack Success Rate), EBD(Empirical Boundary Distance), ACTC(Average Confidence of True Class), ACAC(Average Confidence of Adversarial Class), MP(Minimal Perturbation) and ROBY in terms of independence, effec-

收稿日期:2024-10-16;修回日期:2025-03-15;责任编辑:覃怀银

tiveness, and efficiency, and reduces time consumption by 55% compared to EBD.

Key words: robustness evaluation; adversarial robustness; decision boundary; adversarial attacks; model robustness

Foundation Item(s): National Natural Science Foundation of China (No.62376047, No.62106030); Chongqing Natural Science Foundation Innovation and Development Joint Fund Key Project (No.CSTB2023NSCQ-LZX0003); Chongqing Municipal Education Commission Science and Technology Research Program Key Project (No. KJZD-K202300603); Chongqing Technological Innovation and Application Development Project (No.CSTB2022TIAD-GPX0014); China University Industry-University-Research Innovation Fund Project (No.2022BL105)

1 引言

近年来,深度神经网络(Deep Neural Networks, DNN)在图像识别、语音识别、自然语言处理、计算机视觉和图深度学习等越来越多的研究领域取得了巨大的成功,从而被广泛地应用于地质灾害监测、自动驾驶、蛋白质结构与功能预测、医学图像癌症检测等复杂任务. 然而,Szegedy 等人[1]发现深度学习模型具有对抗脆弱性,容易受到不易察觉的对抗扰动的影响从而导致目标模型输出错误结果,其中被扰动的输入样本被称为对抗样本(Adversarial Examples). 随后,研究人员在文本处理[2]、时序数据预测[3]、图链路预测[4]等领域进一步证实了对抗攻击的存在. 这对深度学习模型的应用任务,特别是对高安全性要求的敏感任务,带来了极大的安全隐患[5~7]. 因此,保障和提升深度学习模型的对抗鲁棒性,即系统在受到对抗攻击的情况下保持其性能水平的能力,变得越来越重要.

为了保障和提升模型的对抗鲁棒性,核心任务是实现模型对抗鲁棒性的度量和评估. 然而,目前研究领域关于模型对抗鲁棒性的评估框架尚未完善,相关的评估方式主要可以分为三类. 其中,第一类评估方式是"基于模型准确性的对抗鲁棒性评估"[8]. 此类方法通过对抗攻击来实现,假设在同样攻击条件下目标模型的准确性越高模型越鲁棒. 为了度量模型在对抗攻击条件下的模型准确性,研究人员提出了任务性能指标 CA(Clean Accuracy)、对抗性能指标 AAW(Adversarial Accuracy on White-box attacks)及 ACAC(Average Confidence of Adversarial Class)、防御性能指标 CAV(Classification Accuracy Variance)等. 同时,此类方法认为良好的对抗样本应该尽可能地覆盖模型的各个神经元并且不易被感知,从而提出了神经元覆盖性度量方法 KMNCov(K-Multisection Neuron Coverage)以及不易感知性度量方法 ASS(Average Structural Similarity)、ALDp(Average ℓ_p Distortion)等. 第二类评估方法是"基于样本间距离的对抗鲁棒性评估"[9],此类方法假设对于一个样本点,存在最小的扰动半径,使得该半径空间内的所有样本都可以被正确预测,而大于该半径的空间可能存在使得模型误判的对抗样本. 如果可以计算出目标模型的扰动半径,那么就可以用于模型鲁棒性评估,取值越大表示目标模型对抗鲁棒性越强. 为此,研究领域提出了扰动半径上边界评估方法 DeepFool[10]以及下边界评估方法 CLEVER Score[11]等. 第三类评估方法是"基于模型输出的对抗鲁棒性评估"[8],代表性指标有 NS(Neuron Sensitivity)和 ε-ENI(ε-Empirical Noise Insensitivity),此类方法通过计算正常样本和对应对抗样本在目标模型中输出的差异性来评估模型的对抗鲁棒性,其假设越鲁棒的模型对对抗攻击越不敏感,正常样本与对抗样本应具有越相似的输出.

尽管研究领域关于深度模型对抗鲁棒性评价进行了初步的研究工作,但现有方法在实际应用中面临诸多挑战. 首先,"基于模型准确性的对抗鲁棒性评估方法"需要执行特定的对抗攻击方法实现对抗鲁棒性评估,评估结果依赖选用的对抗攻击方法、对抗扰动程度、对抗攻击与目标模型的关系等因素,而且成本高昂,任何实验设置方面的差异都会导致目标模型对抗鲁棒性的不可比较性. 其次,"基于样本间距离的对抗鲁棒性评估方法"中扰动半径上边界评估方法同样依赖具体的对抗攻击过程,下边界评估方法主要基于理论证明方式,逼近过程的计算复杂度很高. 最后,"基于模型输出的对抗鲁棒性评估方法"需要多次执行目标模型,基于多个输出的差异性的统计量进行评估,依赖对抗样本的生成方法且具有很高的计算代价.

为了克服以上挑战,本文提出一种基于深度学习模型决策边界光滑度的对抗鲁棒性评估指标 DBSE(Decision Boundary Shannon Entropy). 受到对抗脆弱性与模型决策边界高弯曲区域关系[12]的启发,DBSE 假设对抗扰动通常指向决策边界的高弯曲区域,此区域决策边界的轻微变化会导致被扰动样本产生不同的预测结果,决策边界高弯曲区域越多,模型越容易被攻击,从而模型的对抗鲁棒性与模型决策边界整体的光滑度紧密相关. 为了刻画模型的决策边界,本文提出了基于二分搜索策略的样本采样方法以对决策边界进行间接拟合. 在此基础上,利用奇异值分解来提取决策边界采样样本的空间结构信息,奇异值的分布越均匀,决策边界的结构越复杂. 最终,本文提出基于奇异值香农熵的模型对抗鲁棒性评估指标.

2 相关工作

2.1 深度学习模型对抗鲁棒性评估方法

(1)基于模型准确性的对抗鲁棒性评价

为了评价深度学习模型的对抗鲁棒性,最常见的评价方式是度量目标模型在对抗攻击下的模型性能.根据传统的软件测试理论,测试样本的覆盖度是衡量软件是否得到全面测试的关键因素.基于此思想,研究人员提出了一系列度量对抗样本集合对目标模型神经元覆盖度的评估方法,希望通过增加覆盖范围尽可能发现模型存在的缺陷[8].例如,Pei等人[13]将软件工程领域中的测试充分性类比到模型测试中,提出了神经元覆盖率指标NCov(Neuron Coverage)来量化不同测试数据神经元被激活的比例,NCov的值越大表示对抗样本的充分性越高,评估结果的可信度越高.同时,对抗攻击要求基于尽可能小的对抗扰动使模型产生错误的预测结果.为了度量对抗扰动程度,Marchetti等人[14]提出了ALD_p作为不可感知性的度量指标,ALD_p越小表示样本失真越小,对抗扰动越小.另外,研究人员还提出了ASS、PSD(Perturbation Sensitivity Distance)、Wasserstein Distance等指标[9].基于对抗样本,此类评估方法期望目标模型在对抗攻击下尽可能保持稳定的预测性能.另外,基于对抗样本预测置信度的指标ACTC(Average Confidence of True Class)和ACAC也被用于度量模型的对抗鲁棒性[15],其数学表达式如下:

$$\mathrm{ACTC}=\frac{1}{N}\sum_{i=1}^{N}P\left(f(\boldsymbol{x}_i^{\mathrm{adv}})=\boldsymbol{y}_i\right) \tag{1}$$

$$\mathrm{ACAC}=\frac{1}{N}\sum_{i=1}^{N}P\left(f(\boldsymbol{x}_i^{\mathrm{adv}})\neq\boldsymbol{y}_i\right) \tag{2}$$

其中,$\boldsymbol{x}_i^{\mathrm{adv}}$表示对抗样本,$N$表示样本数量,$f(\cdot)$表示模型的输出,$P(\cdot)$表示模型分类的置信度.正确类别平均置信度ACTC表示对抗样本被分类成正确类别时模型置信度的平均值,取值越高,说明模型正确识别对抗样本类别的能力越强,对抗鲁棒性越强.对抗类别平均置信度ACAC为模型将对抗样本进行错误分类的置信度的平均值,取值越高,说明模型识别对抗样本类别的能力越差,对抗鲁棒性越差.此外,攻击成功率ASR(Attack Success Rate)是被广泛接受和使用的鲁棒性评估指标[16].ASR是指对正常样本添加对抗扰动成功欺骗模型的对抗样本与总体样本数量的比值,数学表达式如下:

$$\mathrm{ASR}=\frac{1}{N}\sum_{i=1}^{N}\mathrm{count}\left(f(\boldsymbol{x}_i^{\mathrm{adv}})\neq\boldsymbol{y}_i\right) \tag{3}$$

其中,count(·)是计数函数.为了测评模型的预测性能,AAW指标也被提出以度量对抗样本被模型误分类的比例[16].总体上,基于模型准确性的对抗鲁棒性评估指标可以快速给出直观的评估结果,是一种简单有效的对抗鲁棒性评估方法.但该类指标依赖对抗样本的质量,不同对抗攻击算法生成的对抗样本对评估指标的结果影响较大,不具有唯一性.

(2)基于样本间距离的对抗鲁棒性评价

在基于样本间距离的对抗鲁棒性评价指标中,对抗距离的上边界方法假设对于实际距离大于对抗距离上边界的样本,存在一种扰动使得其可以变为对抗样本,即最小扰动量(Minimal Perturbation,MP)[10].这类方法通常通过设计算法去构造扰动更小的对抗样本来实现,因此大部分是攻击相关的评估方式[9].例如,Moosavi-Dezfooli等人[10]提出用于构造细微扰动的对抗攻击算法DeepFool,迭代过程中第i轮的扰动$\boldsymbol{r}_i$的数学表达式如下:

$$\arg\min_{\boldsymbol{r}_i}\|\boldsymbol{r}_i\|_2 \ \text{ subject to } f(\boldsymbol{x}_i)+\nabla f(\boldsymbol{x}_i)^{\mathrm{T}}\boldsymbol{r}_i=0 \tag{4}$$

不断迭代以上过程,直到样本$\boldsymbol{x}_i^{\mathrm{adv}}$的预测类别发生变化.Bastani等人[17]提出了用于鲁棒性评估的对抗样本构造方法LP formulation.然而,此类方法大部分是与对抗攻击相关的评估方式,本质上仍然没有摆脱对抗样本对评估任务的影响,依然受制于对抗攻击算法.另一方面,对抗距离的下边界方法目标是寻找一个样本间距离的下边界,使得小于该距离的扰动都无法使得原样本被转化为对抗样本.Weng等人[11]将对抗距离的下边界转化为局部Lipschitz常数估计问题,并提出了攻击无关的、适用于任何神经网络的CLEVER和二阶CLEVER指标.另外,Zhang等人[18]提出了针对任意激活函数的通用鲁棒性证明的框架CROWN,从而提供一个更接近真实对抗距离的可证明的下边界.虽然此类方法与攻击无关、通用性强,但计算复杂度很高.

(3)基于模型输出的对抗鲁棒性评价

为了度量模型的对抗鲁棒性,研究人员提出基于目标模型对正常样本和对抗样本的处理结果的差异性进行鲁棒性评价.Zhang等人[19]提出神经元敏感性评估指标NS(Neuron Sensitivity),给定正常样本$\boldsymbol{x}_i$和对应的对抗样本$\boldsymbol{x}_i^{\mathrm{adv}}$,构成样本对集合$\bar{D}=\left\{(\boldsymbol{x}_i,\boldsymbol{x}_i^{\mathrm{adv}})\right\}$,数学表达式如下:

$$\delta(f_l^m,\bar{D})=\frac{1}{N}\sum_{i=1}^{N}\frac{1}{\dim\left(f_l^m(\boldsymbol{x}_i)\right)}\left\|f_l^m(\boldsymbol{x}_i)-f_l^m(\boldsymbol{x}_i^{\mathrm{adv}})\right\|_1 \tag{5}$$

其中,$f_l^m(\boldsymbol{x}_i)$表示神经元对样本$\boldsymbol{x}_i$的输出.另外,研究人员提出了基于正常样本和对抗样本模型损失函数差异性的评估方法ε-ENI[8].然而,此类方法依然需要对抗攻击方法生成对抗样本才能够进行度量,评估结果依赖具体的对抗样本生成机制.

与以上方法不同,本文的目标是提出不依赖对抗攻击方法的、反映目标模型自身内在特征的对抗鲁棒

性评估指标. 和本文目标类似的深度模型对抗鲁棒性评估指标有经验决策边界距离EBD(Empirical Boundary Distance)[20]和ROBY[21]. EBD指标本质上是所有输入样本到决策边界的经验距离平均值,数学表达式如下:

$$\mathrm{EBD}=\frac{1}{N}\sum_{i=1}^{N}d_i=\frac{1}{N}\sum_{i=1}^{N}\min\phi_i(V) \tag{6}$$

其中,V表示互相正交的随机方向集合,$\phi_i(\cdot)$表示样本$\boldsymbol{x}_i$在特定方向上到决策边界的均方根距离,d_i表示模型关于样本$\boldsymbol{x}_i$输出结果发生变化的最小距离. 对抗鲁棒性评估指标ROBY使用类内特征子空间聚合(Feature Subspace Aggregation, FSA)和类间特征子空间距离(Feature Sub-space Distance,FSD)来衡量模型面对对抗攻击的鲁棒性,其定义如下:

$$\mathrm{ROBY}=\frac{\sum_{i=1}^{K-1}\sum_{j=i+1}^{K}\left(\mathrm{FSA}_i+\mathrm{FSA}_j-\mathrm{FSD}_{i,j}\right)}{K(K-1)/2} \tag{7}$$

其中,K表示所有分类的类别总数.

2.2　基于决策边界的模型对抗鲁棒性研究

在深度学习模型对抗鲁棒性分析方面,为了深化对模型决策机制的理解,Karimi等人[22]提出了基于对抗样本的深度学习模型决策边界附近样本生成方法DeepDIG,从而基于此分析模型决策边界的特性. Fawzi等人[23]认为分类器的鲁棒性与决策边界的几何形状密切相关,分析了深度学习模型面对扰动的鲁棒性,并在此基础上对这类分类器的复杂决策边界的几何特性进行了探索研究,从而支撑模型架构的优化改进. He等人[24]在模型扰动输入条件下对决策边界的表达能力进行了研究,结果表明扰动条件下的决策边界与正常样本周围的决策边界具有不同特征. Yu等人[25]通过模型损失函数的可视化解释对抗攻击和防御机制,认为损失函数关于模型输入的曲面越平滑模型越鲁棒,并通过利用KL距离(Kullback-Leibler divergence)度量不同样本预测结果之间的差异性设计了损失函数曲面平滑性分析方法. 为了从决策边界的角度理解深度神经网络的泛化性,从模型优化算法以及训练数据角度出发,Lei等人[26]提出了算法决策边界变化性(Algorithm Decision Boundary Variability)和数据决策边界变化性(Data Decision Boundary Variability)概念,通过实证分析发现决策边界的变化性与模型泛化性之间存在显著的负相关性.

在深度学习模型对抗鲁棒性优化方面,Yan等人[27]基于高维决策空间中正常样本及其对抗样本的测地距离定义了测地损失函数,并以此作为正则化项来捕获自然样本和对抗样本之间的最小分布偏移,从而提出了测地对抗训练方法GeodesicAT. 模型决策边界的复杂度反映了训练分布的复杂度和模型学习的难度,它与训练样本之间的间隔反映了模型的鲁棒性. Chen等人[28]面向对抗训练提出了一个决策边界感知数据增强框架CODA. 在每轮训练过程中,CODA直接使用前一轮的元信息来指导增强过程,生成更多接近决策边界的数据,即攻击数据. Kanbak等人[29]研究了深度学习模型面对图像几何变换的脆弱性,提出了攻击方法ManiFool以通过测地距离获得一个小的最坏情况几何变换并计算分类器的不变性分数,在此基础上使用几何变换执行对抗性训练从而提升模型的鲁棒性.

本文以模型决策边界与模型预测之间的关系为基础,利用已有的基于决策边界的模型对抗鲁棒性的相关认识和理解,提出能够准确度量深度学习模型对抗鲁棒性的评估方法.

3　方法描述

本节将详细介绍提出的基于模型决策边界的对抗鲁棒性评估指标DBSE. DBSE由决策边界样本采样和对抗鲁棒性指标设计两部分构成,其中决策边界样本采样方法利用决策边界的数学定义对不同类别的样本进行二分搜索以获取“决策边界样本”,从而近似决策边界的分布情况和结构特征. 为了度量模型决策边界的光滑度,评估指标DBSE先利用奇异值分解提取决策边界样本集的特征信息,再经过香农熵计算得到最后的评估指标取值. DBSE没有对抗样本的参与,从根本上避免了传统对抗鲁棒性评估对对抗攻击方法的依赖性,并且DBSE是一个黑盒的评估过程,更加符合现实应用场景. 具体框架如图1所示.

3.1　决策边界样本采样方法

基于相同的训练数据,深度学习模型采用不同的模型架构和训练策略会导致模型形成不同的决策边界. 大量的研究表明,深度学习模型的决策边界与模型的对抗鲁棒性密切相关[22]. 然而,高维决策空间中的模型决策边界难以被直接观测,很难获得模型决策边界的准确描述. 因此,本文提出利用模型决策边界附近的样本集合来近似其在高维空间中的分布情况及结构信息.

为了获得深度学习模型决策边界附近的样本集合,对抗样本对正常样本添加精心设计的微小扰动以使其恰好跨越模型决策边界,它们一般被认为是距离模型决策边界较近的数据点,因此本文可以基于此特性进行样本选择从而近似刻画模型决策边界. 然而,基于对抗样本的刻画方法容易受到对抗攻击方法和扰动程度等因素的影响. 同时,对于采用了防御措施的鲁棒性模型,攻击成功率会明显下降甚至无法成功,并且对抗攻击的开销成本也会随之增加. 因此,利用对抗攻击

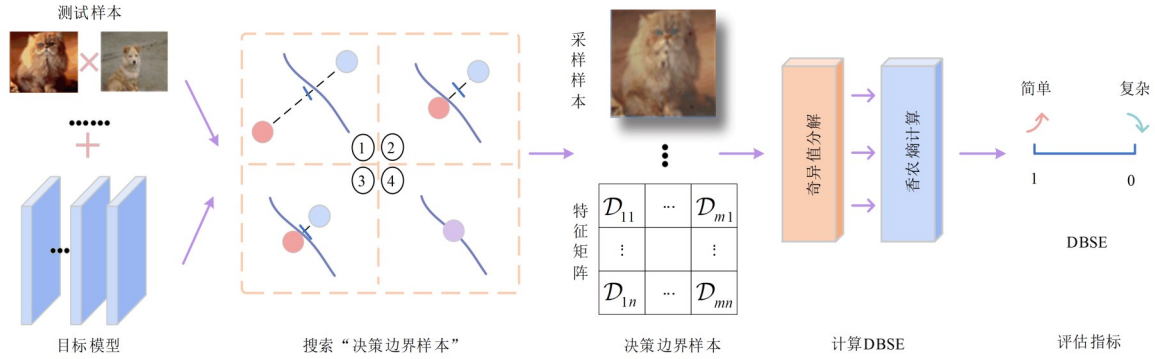


图1 对抗鲁棒性评估指标DBSE框架图

获取对抗样本以近似模型决策边界的方法无法直接满足实际需求. 在此基础上,本文结合模型决策边界的基本概念,提出通过直接在模型的正常输入样本中搜索决策边界附近样本以近似模型决策边界的思路.

给定一个面向C个类别的多分类深度学习模型$f(\cdot)$,对于数据集$\boldsymbol{X}$中的输入样本$\boldsymbol{x}\in R^{n\times n}$,令其关于第$i$类的分类置信度和第$j$类的分类置信度分别为$f_i(\boldsymbol{x})$和$f_j(\boldsymbol{x})$,同时令模型在$C$个类别中的最大置信度为$\max_{k\in C}f_k(\boldsymbol{x})$. 如果$f_i(\boldsymbol{x})$和$f_j(\boldsymbol{x})$相等且等于$\max_{k\in C}f_k(\boldsymbol{x})$,那么可以认为样本$\boldsymbol{x}$位于深度学习模型的决策边界上,形式化定义如下:

$$\mathrm{DB}=\left\{\boldsymbol{x}\in\boldsymbol{X}\mid\exists i,j\in C,\ i\neq j,\ f_i(\boldsymbol{x})=f_j(\boldsymbol{x})=\max_{k\in C}f_k(\boldsymbol{x})\right\} \tag{8}$$

然而,对于实际模型而言,往往很难出现刚好完全位于决策边界上的样本,即往往难以$f_i(\boldsymbol{x})=f_j(\boldsymbol{x})=\max_{k\in C}f_k(\boldsymbol{x})$的要求. 为此,本文通过设置阈值$\gamma$对以上约束条件进行放松,对关于两个类别的置信度差值小于此阈值的样本$\boldsymbol{x}$,认为其位于决策边界附近,可以用于近似刻画决策边界的分布和结构信息,形式化定义如下:

$$\mathrm{DB}=\left\{\boldsymbol{x}\in\boldsymbol{X}\mid\exists i,j\in C,\ i\neq j,\ \left|f_i(\boldsymbol{x})-f_j(\boldsymbol{x})\right|\leqslant\gamma,\ f_j(\boldsymbol{x})=\max_{k\in C}f_k(\boldsymbol{x})\right\} \tag{9}$$

基于以上决策边界的定义,本文提出基于二分搜索的深度学习模型决策边界样本采样算法,具体如算法1所示. 首先从数据集$\boldsymbol{X}$中随机选择两个具有不同类别的样本$\boldsymbol{x}$和$\boldsymbol{y}$,如果二分样本$\boldsymbol{z}$满足决策边界条件且不属于其他类别,则更新决策边界样本集合DB. 否则,更新样本$\boldsymbol{x}$或$\boldsymbol{y}$. 如果样本$\boldsymbol{z}$与$\boldsymbol{x}$和$\boldsymbol{y}$的类别都不一致,则不属于决策边界,直接舍弃,重新开始搜索.

算法1 决策边界样本采样算法

输入:目标模型$f(\cdot)$,数据集$\boldsymbol{X}$,阈值γ,采样数量K

输出:决策边界样本集合DB

1. DB=∅, 数量$k=0$ /*初始化*/
2. WHILE $\exists\boldsymbol{x},\boldsymbol{y}\in\boldsymbol{X}$ & $c_x\neq c_y$ & $k\leqslant K$ /*判断是否满足采样数量*/
3. $\boldsymbol{z}=(\boldsymbol{x}+\boldsymbol{y})/2$ /*计算决策边界样本 */
4. IF $\left|f_i(\boldsymbol{z})-f_j(\boldsymbol{z})\right|\leqslant\gamma$&$f(\boldsymbol{z})\in\{c_x,c_y\}$ /*判断样本$\boldsymbol{z}$是否满足条件*/
5. $k=k+1$, DB=DB$\cup\{\boldsymbol{z}\}$ /*将样本$\boldsymbol{z}$加入决策边界样本集*/
6. break
7. ELSE
8. IF $f(\boldsymbol{z})=c_x$ /*判断样本$\boldsymbol{z}$是否与$\boldsymbol{x}$的类别一致*/
9. $\boldsymbol{x}=\boldsymbol{z}$
10. ELSE IF $f(\boldsymbol{z})=c_y$ /*判断样本$\boldsymbol{z}$是否与$\boldsymbol{y}$的类别一致*/
11. $\boldsymbol{y}=\boldsymbol{z}$
12. ELSE
13. break /*样本$\boldsymbol{z}$的类别与$\boldsymbol{x}$,$\boldsymbol{y}$都不一致,穿过了其他区域,重新开始*/
14. END IF
15. END IF
16. END WHILE
17. RETURN决策边界样本集合DB

3.2 对抗鲁棒性评估指标

本文假设模型对抗鲁棒性与决策边界整体的光滑度紧密相关,具有越光滑决策边界的模型往往越鲁棒. 虽然采样得到的决策边界样本集合DB可以近似模型决策边界的分布情况和结构特征,但依然无法提供合适的信息来量化决策边界的光滑度. 为此,本文提出结合奇异值分解和香农熵来度量模型决策边界的光滑度.

图2显示了对抗鲁棒性评估指标DBSE的原理图. 其中,对决策边界样本特征矩阵使用奇异值分解(Singular Value Decomposition),得到左奇异向量矩阵、奇异值矩阵和右奇异向量矩阵,这些奇异向量表示了决策边界的主要变化方向,而奇异值则是这些变化方向的权重. 奇异值越大,表示数据在该方向上的变化越显著,从而这个方向就越重要. 因此,如果这些奇异值的分布越均匀,那么表示每个方向对于决策边界都很重要,决策边界的结构越复杂. 反之,如果奇异值的分布越集中,那么决策边界只会集中在少数几个变化方向,

决策边界的结构相对较为简单. 本文利用这一规律评估模型的对抗鲁棒性.

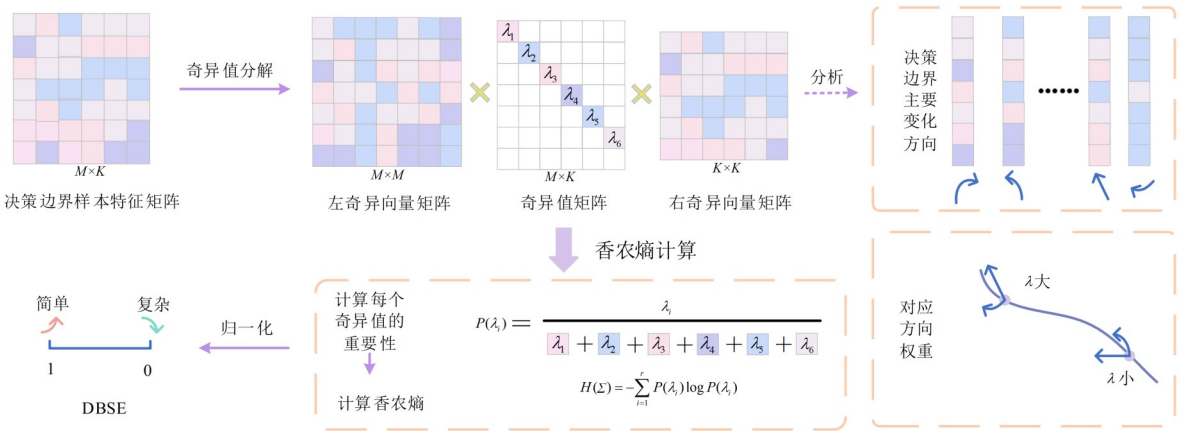


图2　评估指标DBSE的计算原理图

对于决策边界样本集合DB,首先获得目标模型中各个决策边界样本$\boldsymbol{z},\boldsymbol{z}\in\mathrm{DB}$,对应的嵌入表示,并将各个嵌入表示进行拼接形成决策边界样本特征矩阵$\boldsymbol{M}_X$,$\boldsymbol{M}_X\in\boldsymbol{R}^{M\times K}$. 然后,对矩阵$\boldsymbol{M}_X$进行奇异值分解,获得奇异向量和奇异值矩阵,确定决策边界的主要变化方向及各个变化方向上的权重,形式化定义如下:

$$\begin{aligned}\boldsymbol{M}_X&=\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^{\mathrm{T}}\\&=\begin{pmatrix}u_{11}&&u_{1M}\\\vdots&\ddots&\vdots\\u_{M1}&&u_{MM}\end{pmatrix}_{M\times M}\cdot\begin{pmatrix}\lambda_1&&&&\\&\ddots&&&\\&&\lambda_r&&\\&&&\ddots&\\&&&&0\end{pmatrix}_{M\times K}\\&\quad\cdot\begin{pmatrix}v_{11}&&v_{1K}\\\vdots&\ddots&\vdots\\v_{K1}&&v_{KK}\end{pmatrix}^{\mathrm{T}}_{K\times K}\end{aligned}\tag{10}$$

其中,$\boldsymbol{U}$是矩阵$\boldsymbol{M}_X$的左奇异向量矩阵,$\boldsymbol{\Sigma}$是奇异值矩阵,$\boldsymbol{V}$是右奇异值向量矩阵. 奇异向量矩阵中的列向量表示决策边界特征的变化方向,奇异值$\{\lambda_1,\lambda_2,\cdots,\lambda_r\}$则是这些主要变化方向的权重.

香农熵(Shannon Entropy)可用于量化信息的不确定性或信息的混乱程度. 对于均匀分布的数据,数值的不确定性最大,香农熵达到最大值;对于非均匀分布的数据,即某些数值出现的概率明显高于其他数值时,香农熵的值较小. 本文采用香农熵度量奇异值分解中各个变化方向的分布情况,将奇异值的占比视为数值的出现概率:

$$P(\lambda_i)=\frac{\lambda_i}{\sum_{j=1}^{r}\lambda_j}\tag{11}$$

$P(\lambda_i)$反映了奇异值λ_i在所有奇异值中的重要性. 从而,利用香农熵度量决策边界光滑度的定义如下:

$$H(\Sigma)=-\sum_{i=1}^{r}P(\lambda_i)\log P(\lambda_i)\tag{12}$$

其中,$H(\Sigma)$表示奇异值$\Sigma=\{\lambda_1,\lambda_2,\cdots,\lambda_r\}$的熵,$r$表示奇异值的总数. 最后将$H(\Sigma)$进行归一化就可以得到最终的对抗鲁棒性评估指标DBSE:

$$\mathrm{DBSE}=1-\frac{-\sum_{i=1}^{r}P(\lambda_i)\log P(\lambda_i)}{\log(r)}\tag{13}$$

DBSE值越大,表示决策边界的空间结构越简单、越光滑,模型对抗鲁棒性越强. 反之,DBSE的值越小,表示模型决策边界的空间结构越复杂,模型的对抗鲁棒性越弱.

4　实验及结果分析

为了评估对抗鲁棒性评估指标DBSE的性能,本文将其与6种代表性的模型鲁棒性评估指标ASR[16]、EBD[20]、ACTC[15]、ACAC[15]、MP[10]和ROBY[21]进行对比分析. 为了方便评价,本文定义ACA(ACA=1−ASR)、ACAC′(ACAC′=1−ACAC)和ROBY′(ROBY′=1−ROBY)代替ASR、ACAC和ROBY进行鲁棒性评价,使得指标取值越大表示目标模型越鲁棒. 此外,本文还分析了指标DBSE在不同鲁棒性程度的模型上的取值变化情况以及与代表性评估指标的相关性. 最后,对比分析了各个鲁棒性评估指标的时间效率,探究了参数取值对评估指标DBSE的影响.

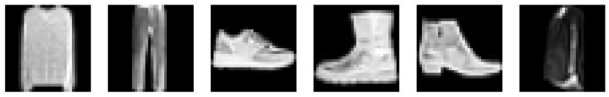
4.1　实验设置

实验评价采用图像领域经典数据集CIFAR-10、MNIST和Fashion MNIST. 其中,CIFAR-10是一个彩色图像识别数据集,它包含10个类别,有50 000个训练图片和10 000个测试图片,全部为32×32像素的RGB彩色图片;MNIST是手写数字识别数据集,它包含10个类别,代表数字0~9,

其中有60 000个训练图片和10 000个测试图片,全部为28×28像素的灰度图片;Fashion MNIST数据集由背景为浅灰色、分辨率为762×1 000的JPEG原始图片经过处理得到,包含10个衣服类别,其中有60 000个训练图片,10 000个测试图片,全部为28×28像素的灰度图片. 各数据集样本示例如图3所示.

(a) CIFAR-10

(b) MNIST



(c) Fashion MNIST

图3 各数据集的样本示例

实验采用的环境为RTX 3050 Ti Laptop GPU和Pytorch深度学习框架. 本文选用Resnet34、Resnet50、Resnet101、Alexnet、Squeezenet、Densenet、Googlenet、Mobilenet、Shufflenet共9种基于深度学习的图像分类模型,并利用对抗训练来获取其具有不同鲁棒性程度的模型. 在相关参数取值方面,ACA指标基于最强一阶攻击算法PGD(Projected Gradient Descent),迭代次数为10,最大扰动为8/255,扰动步长设置为1/255;EBD指标选取150个正常样本计算经验边界距离;DBSE指标每个决策边界集合采样数K设置为175,置信度阈值γ设置为0.01.

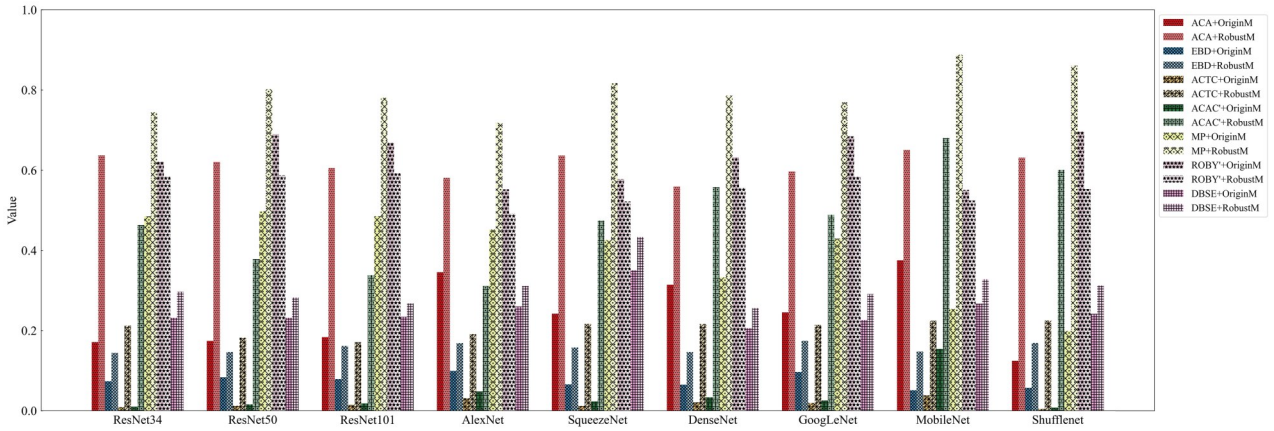
4.2 有效性验证

为了验证对抗鲁棒性评估指标DBSE的有效性,本文将其与评估指标ACA、EBD、ACTC、ACAC′、MP、ROBY′进行对比分析,其中MP使用的对抗攻击算法是DeepFool. 为了构建鲁棒性模型,实验中对抗训练使用的对抗样本生成算法是FGSM(Fast Gradient Sign Method)[30],扰动大小设置为4/255. 实验结果如图4所示.

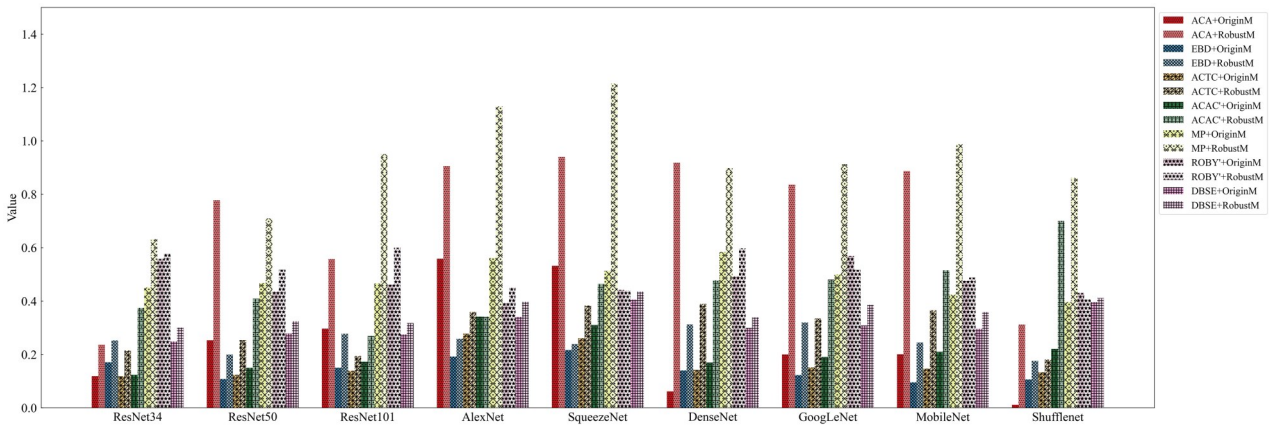
图4(a)~图4(c)分别显示了在CIFAR-10、MNIST、Fashion MNIST数据集上,9种模型的原始版本(OriginM)和经过对抗训练后形成的鲁棒性版本(RobustM)对应的各指标的取值. 结果显示,ACA、EBD、ACTC、ACAC′和MP指标在对抗训练后的鲁棒性模型上的取值明显高于在原始模型上的取值,即ACA+RobustM、EBD+RobustM、ACTC+RobustM、ACAC′+RobustM和MP+RobustM的值明显高于ACA+OriginM、EBD+OriginM、ACTC+OriginM、ACAC′+OriginM和MP+OriginM,这表明了上述指标确实能够度量模型的对抗鲁棒性. 同时,实验结果表明,DBSE指标在对抗训练后的模型上的表现(DBSE+RobustM)同样优于在各个原始模型上的表现(DBSE+OriginM),这与以上指标的结果一致,说明了本文提出的对抗鲁棒性评估指标DBSE能够有效区分不同模型的对抗鲁棒性. 然而,需要注意的是,ROBY′指标的表现与预期不符,在鲁棒性模型上表现出更低的取值,因此本文后续有效性对比实验中将不再使用ROBY′作为对比方法.

为了进一步充分验证DBSE指标在模型对抗鲁棒性变化过程中的稳定性和有效性,本文比较了ACA、EBD、ACTC、ACAC′、MP和DBSE指标对具有不同鲁棒性程度的模型的评估结果. 为了基于对抗训练构建鲁棒性模型,本实验在对抗样本生成算法FGSM的基础上,增加了BIM(Basic Iterative Method)[31]和PGD(Projected Gradient Descent)[32]方法. 另外,本文从上述的九种模型中选取了ResNet34、AlexNet和SqueezeNet三个模型进行进一步的实验. 具体的,从残差网络(ResNet)系列模型中,包括ResNet34、ResNet50和ResNet101,以及特征复用架构的DenseNet等模型中选择ResNet34模型. ResNet34虽然层数相对较少,但其残差结构具有代表性,能够有效解决深层网络中的梯度消失问题;由于AlexNet首次引入Dropout技术并大规模使用ReLU激活函数,并实证了深度卷积网络在大规模图像分类任务中的有效性,为后续深度学习的发展奠定了基础,因此本文在传统卷积神经网络中选择AlexNet模型;由于SqueezeNet首创了系统化的轻量化设计方法,为轻量化模型的设计提供了重要的参考和借鉴,因此本文在轻量级网络中选择SqueezeNet模型.

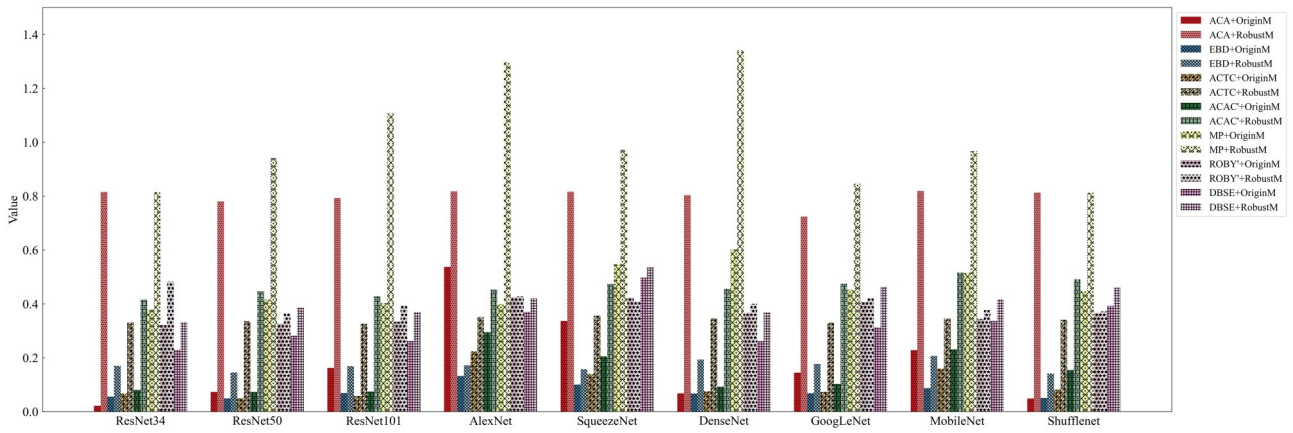
图5展示了基于采用不同参数设置的对抗样本生成算法FGSM进行对抗训练后,具有不同对抗鲁棒性程度的ResNet34、AlexNet和SqueezeNet模型上各个鲁棒性评估指标的变化情况. 在实验中,FGSM的最大扰动幅度分别设置为0、1/255、2/255、…、6/255. 实验结果表明,随着模型对抗鲁棒性的增强,ACA、EBD、ACTC、ACAC′、MP和DBSE的评估结果均呈现出相应的上升趋势. 这一现象验证了DBSE指标能够随着模型对抗鲁棒性的提升而增加,从而持续、有效地反映模型的鲁棒性变化. 为了验证DBSE指标在不同对抗样本生成算法下的表现,图6和图7分别展示了使用BIM和PGD生成对抗样本进行对抗训练后,具有不同对抗鲁棒性程度的ResNet34、AlexNet和SqueezeNet模型上各鲁棒性评估指标的变化情况. 在实验结果中,同样观察到了与图5一致的趋势,即随着模型对抗鲁棒性的增强,ACA、EBD、ACTC、ACAC′、MP和DBSE六个指标的评估值总体上不断上升. 这些实验结果进一步证实了DBSE



(*a*) CIFAR-10数据集各鲁棒性评估指标结果对比图(取值越大模型越鲁棒)



(*b*) MNIST数据集各鲁棒性评估指标结果对比图(取值越大模型越鲁棒)



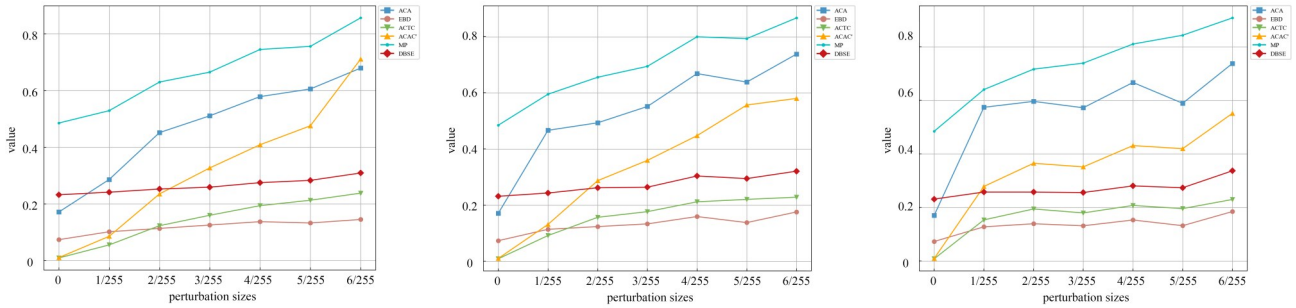
(*c*) Fashion MNIST数据集各鲁棒性评估指标结果对比图(取值越大模型越鲁棒)

图4　CIFAR-10、MNIST、Fashion MNIST数据集上ACA、EBD、ROBY′、DBSE等指标的评估结果

指标在评估模型对抗鲁棒性方面的有效性. 另外,通过观察图5、图6和图7中的实验结果可知,随着模型鲁棒性的不断提升,ACA、ACAC′和MP指标具有一定的波动性,特别是在图7基于PGD方法的实验中,ACA指标取值甚至出现局部下降. 相对地,本文所提DBSE指标在模型对抗鲁棒性不断变化的情况下表现出了良好的稳定性,能提供准确的评估结果

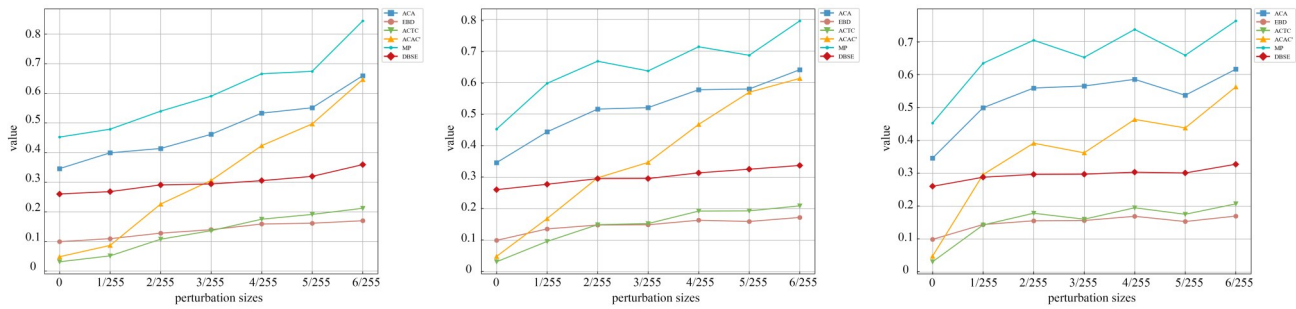
4.3　相关性分析

为了进一步验证DBSE指标与模型的对抗鲁棒性之间的强相关性,本文对具有不同对抗鲁棒性程度的模型进行了评估. 具体来说,本文分析了DBSE和ACA、EBD、ACTC、ACAC′、MP指标在多个不同鲁棒性



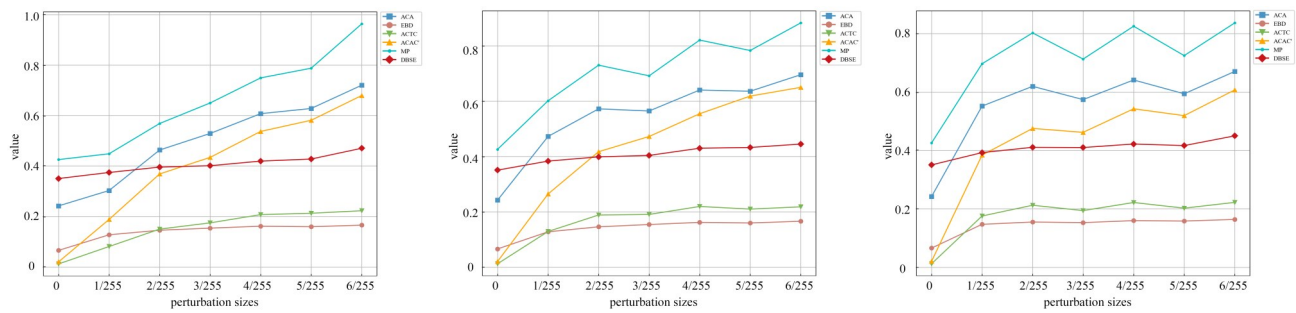
(a) ResNet34 (b) AlexNet (c) SqueezeNet

图5 基于FGSM的对抗训练条件下各模型的ACA、EBD、ACTC、ACAC′、MP和DBSE指标的变化



(a) ResNet34 (b) AlexNet (c) SqueezeNet

图6 基于BIM的对抗训练条件下各模型的ACA、EBD、ACTC、ACAC′、MP和DBSE指标的变化



(a) ResNet34 (b) AlexNet (c) SqueezeNet

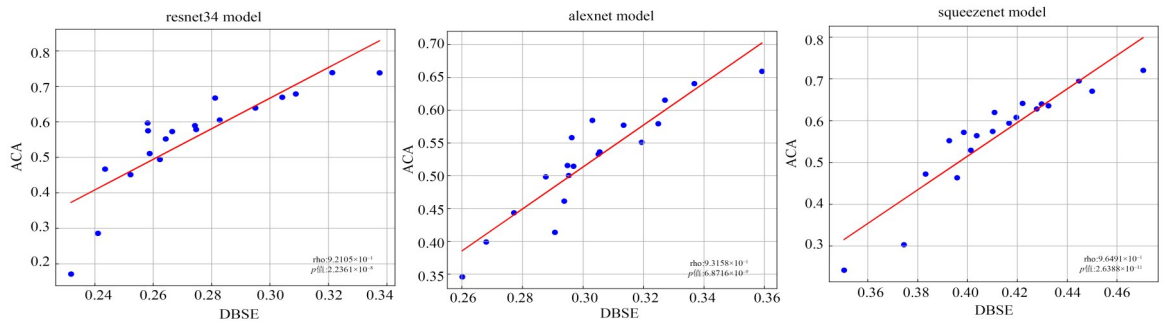
图7 基于PGD的对抗训练条件下各模型的ACA、EBD、ACTC、ACAC′、MP和DBSE指标的变化

模型上的评估结果. 图8展示了ResNet34、AlexNet和SqueezeNet模型上DBSE与各指标的评估结果以及斯皮尔曼(Spearman)秩相关系数,计算公式如下:

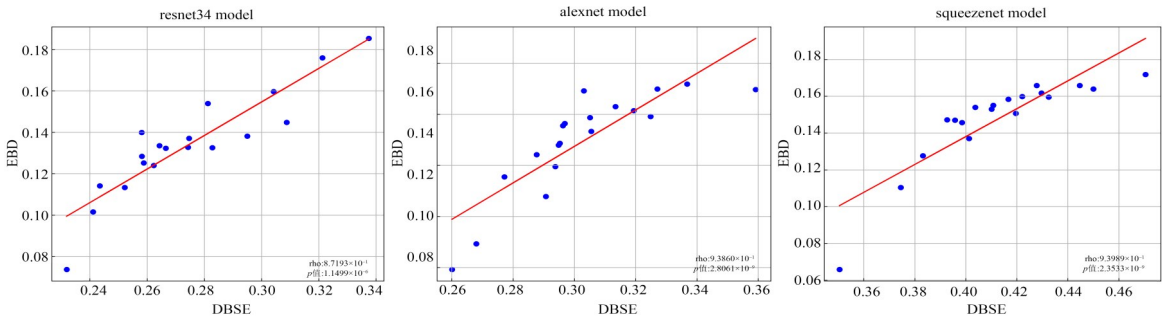
$$\mathrm{rho}=\frac{\frac{1}{n}\sum_{i=1}^{n}\left(R(x_i)-\overline{R(x)}\right)\cdot\left(R(y_i)-\overline{R(y)}\right)}{\sqrt{\left(\frac{1}{n}\sum_{i=1}^{n}\left(R(x_i)-\overline{R(x)}\right)^2\right)\cdot\left(\frac{1}{n}\sum_{i=1}^{n}\left(R(y_i)-\overline{R(y)}\right)^2\right)}} \tag{14}$$

其中,rho表示斯皮尔曼秩相关系数,$R(x)$、$R(y)$分别表示x、y的位次,$\overline{R(x)}$、$\overline{R(y)}$分别表示平均位次. 斯皮尔曼秩相关系数的计算还包括一个p值,用于评估观测到的相关性是否具有统计学意义.

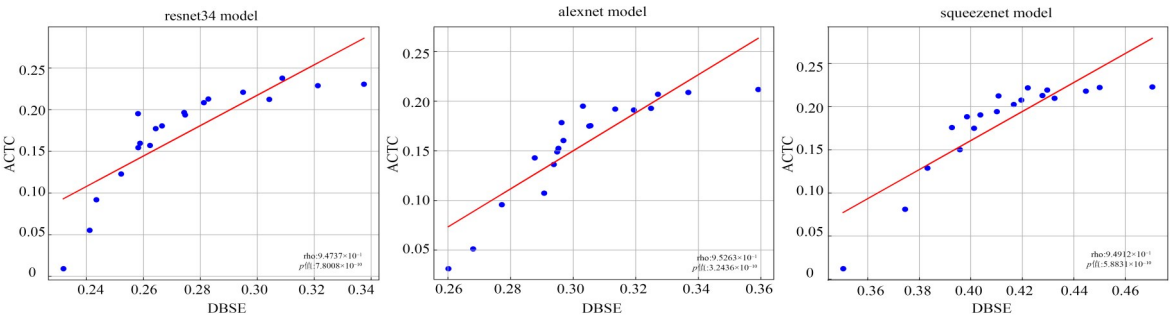
图8(a)~图8(e)分别展示了指标DBSE与指标ACA、EBD、ACTC、ACAC′和MP在不同模型上的评估结果的散点图和斯皮尔曼秩相关系数. 通过观察可以发现,指标DBSE与ACA、EBD、ACTC、ACAC′和MP指标的斯皮尔曼秩相关系数均超过0.85且p值小于10^{-5},即指标DBSE与ACA、EBD、ACTC、ACAC′和MP均存在显著的正相关关系. 由于已知指标ACA、EBD、ACTC、ACAC′和MP与模型对抗鲁棒性之间存在强正相关性,因此可推断出DBSE与对抗鲁棒性之间也存在正相关关系. 以上实验结果表明,DBSE指标能够有效地评估模型对抗鲁棒性.



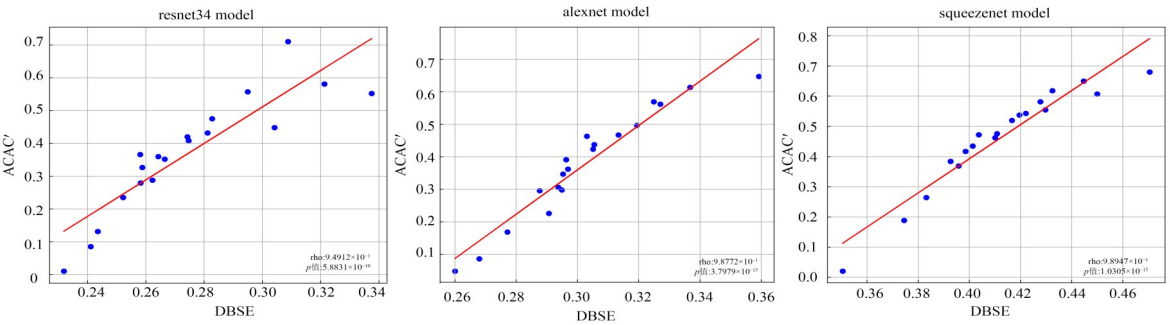
(*a*) DBSE 和 ACA 评估结果散点图与相关性分析



(*b*) DBSE 和 EBD 评估结果散点图与相关性分析

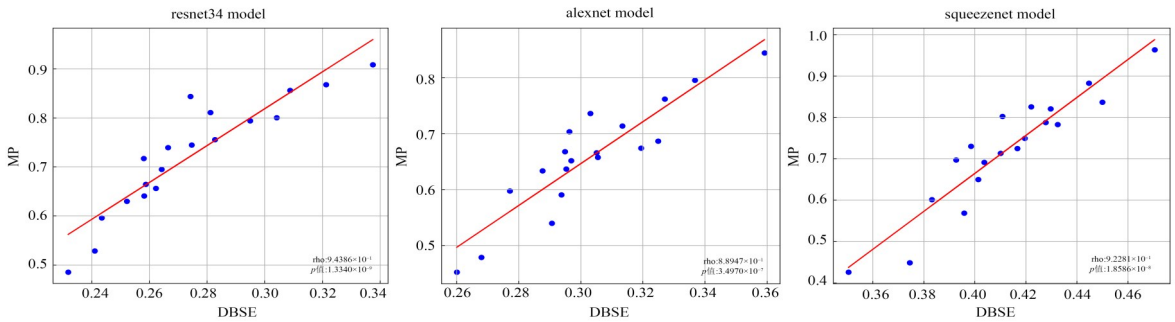


(*c*) DBSE 和 ACTC 评估结果散点图与相关性分析



(*d*) DBSE 和 ACAC′评估结果散点图与相关性分析

图 8　评估指标 DBSE 和 ACA、EBD、ACTC、ACAC′、MP 评估结果散点图与相关性分析



(*e*) DBSE和MP评估结果散点图与相关性分析

续图8

以上实验通过DBSE与其他鲁棒性评估指标的正相关性证明了DBSE的有效性,但其他鲁棒性评估指标并不是客观的评价标准. 因此,为了进一步分析DBSE的有效性,基于对抗攻击方法FGSM、BIM和PGD,表1给出了DBSE与不同鲁棒性程度的模型对应的扰动量大小的相关性. 在正常情况下,对抗攻击的扰动量越大,模型的鲁棒性越低,DBSE的值应该越小. 根据表1的实验结果,各试验场景下的相关系数为负值,表明DBSE与模型对应的扰动量大小呈负相关性,说明了DBSE指标对模型鲁棒性评估的有效性.

表1 DBSE与模型扰动量的相关性

相关性系数	ResNet34	AlexNet	SqueezeNet
FGSM	−1	−1	−1
BIM	−0.964 3	−1	−1
PGD	−0.831 4	−0.964 2	−0.928 5

4.4 时间效率分析

本文提出的对抗鲁棒性评估指标DBSE具备与攻击无关的特性,有效排除了对抗攻击的影响以及相应的计算成本. 尽管如此,DBSE在搜索决策边界样本的过程中仍需消耗大量资源,这无疑增加了评估过程的时间成本. 但与EBD等需要搜索最小扰动代价的评估指标相比,DBSE在计算效率方面仍显示出一定的优势. 表2展示了在三个不同数据集上使用ResNet34模型时,DBSE与ACA、EBD、ACTC、ACAC′、MP和ROBY′各评估指标的时间开销对比. 具体来说,DBSE在三个数据集上的平均时间开销为961.77 s. 相比之下,ACA和ROBY的平均时间开销分别为75.05 s和0.86 s,ACTC和ACAC′的平均时间开销分别为171.23 s和173.34 s,而EBD和MP的时间开销分别为2 095.69 s和1 448.73 s. 虽然与ACA、ROBY′、ACTC、ACAC′相比,DBSE的时间开销较高,但与EBD和MP相比,DBSE的时间开销分别减少了约55%和34%. 这一结果表明,尽管DBSE需要较大时间开销,但相对于EBD和MP等高开销类评估指标,其仍具有显著的效率优势.

4.5 参数敏感性分析

(1)决策边界采样数K影响分析

本文决策边界样本的确定是基于模型对样本的类别置信度. 理论上,当两个类别的置信度差值(即阈值)为0时,决策边界样本精确地位于真实决策边界上. 然而,实际应用中为了获得以上样本,需要大量的输入样例和计算资源. 因此,本文引入了阈值γ,用于搜索接近决策边界的样本来近似地刻画模型的真实决策边界. 为了验证不同阈值对评估指标DBSE的影响,在CIFAR-10数据集上,使用ResNet34模型,对从0.000 1~0.1范围内选取的7个阈值进行了对比实验,结果如表3所示. 从实验结果可知,DBSE指标对阈值的选择并不敏感. 这一现象可以归因于阈值定义的是置信度差值的最大值,而实验中观察到的平均置信度差值均在10^{-2}级数及以下,表明大多数样本的置信度差值确实非常小. 因此,阈值选择应尽可能确保决策边界样本更接近真实决策边界,同时考虑到计算效率,本文选择了0.01作为阈值,这一选择旨在确保评估准确性的同时,最小化计算资源的消耗.

(2)决策边界样本置信度阈值γ影响分析

DBSE指标依赖于模型决策边界附近的输入样本来近似刻画模型的真实决策边界. 因此,采样数K与评

表2 不同指标的时间开销 单位:s

	ACA	EBD	ACTC	ACAC′	MP	ROBY′	DBSE
CIFAR-10	74.54	2 014.26	169.83	173.49	1 416.83	0.83	893.25
MNIST	74.76	2 136.65	172.37	175.16	1 448.64	0.89	905.62
Fashion MNIST	75.84	2 136.15	171.48	171.36	1 480.72	0.86	1 086.43
平均时间开销	75.05	2 095.69	171.23	173.34	1 448.73	0.86	961.77

表3　阈值γ对指标的影响

γ	0.000 1	0.000 5	0.001	0.005	0.01	0.05	0.1 ∞
DBSE	0.226 4	0.227 9	0.228 4	0.227 6	0.228 1	0.221 8	0.224 7
平均置信度差值	4.9×10^{-5}	0.000 2	0.000 4	0.002 3	0.004 5	0.022 1	0.044 4
时间开销/s	2 970.78	1 294.75	1 078.66	973.33	893.25	760.78	685.29

估结果具有相关性. 图9展示了在决策边界样本的采样数K不同取值条件下DBSE指标取值的变化情况. 根据实验结果可知,在采样数较小的情况下(25~125之间),由于采样数不足,无法充分刻画模型决策边界的结构信息,导致DBSE指标不稳定. 这表明在采样数较少时,评估结果不够准确. 随着采样数的增加(150之后),DBSE指标开始趋于稳定. 这表明增加采样数有助于更全面地获得模型决策边界的信息. 考虑到计算成本和效率问题,本文将决策边界采样数K设置为175,以在确保有效地评估模型对抗鲁棒性的同时,避免了过高的计算成本.

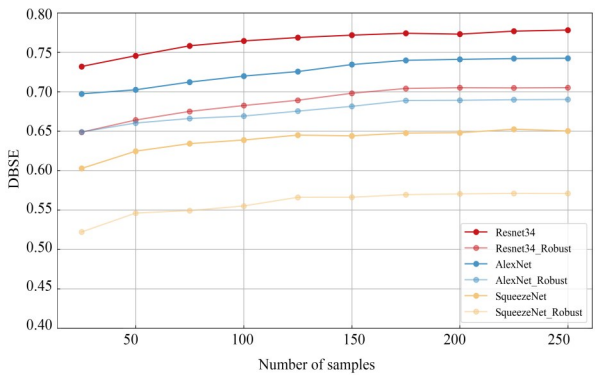


图9　不同采样参数设置下指标DBSE评估结果

5　结论

本文提出了一种用于深度学习模型对抗鲁棒性评估的指标DBSE,该指标通过采样近似刻画模型的真实决策边界,并通过分析度量模型决策边界的光滑度进行模型对抗鲁棒性的评估. 实验结果表明DBSE指标能够有效量化模型的对抗鲁棒性,克服了传统对抗鲁棒性评估指标对对抗攻击的依赖性,并且具有良好的稳定性和时效性. 未来将在进一步深入分析、理解鲁棒性模型决策空间特征性质的基础上,对鲁棒性评估指标的有效性和时效性进行优化.

参考文献

[1] SZEGEDY C. Intriguing properties of neural networks [C]// Proceedings of the International Conference on Learning Representations. Banff: ICLR, 2014: 1-10.

[2] LIU J, JIN H Y, XU G X, et al. Aliasing black box adversarial attack with joint self-attention distribution and confidence probability[J]. Expert Systems with Applications, 2023, 214: 119110.

[3] WU T, WANG X C, QIAO S J, et al. Small perturbations are enough: Adversarial attacks on time series prediction[J]. Information Sciences, 2022, 587: 794-812.

[4] XIAN X P, WU T, QIAO S J, et al. DeepEC: Adversarial attacks against graph structure prediction models[J]. Neurocomputing, 2021, 437: 168-185.

[5] 李明慧, 江沛佩, 王骞, 等. 针对深度学习模型的对抗性攻击与防御[J]. 计算机研究与发展, 2021, 58(5): 909-926.
LI M H, JIANG P P, WANG Q, et al. Adversarial attacks and defenses for deep learning models[J]. Journal of Computer Research and Development, 2021, 58(5): 909-926. (in Chinese)

[6] 吴翼腾, 刘伟, 于淑乔. 基于参数差异假设的图卷积网络对抗性攻击[J]. 电子学报, 2023, 51(2): 330-341.
WU Y T, LIU W, YU X Q. Adversarial attacks on graph convolution networks based on parameter discrepancy hypothesis[J]. Acta Electronica Sinica, 2023, 51(2): 330-341. (in Chinese)

[7] 周侠, 张剑, 李宁安. 基于显著图的电磁信号对抗样本生成方法[J]. 电子学报, 2023, 51(7): 1917-1928.
ZHOU X, ZHANG J, LI N A. An electromagnetic signal adversarial examples generation method based on saliency map[J]. Acta Electronica Sinica, 2023, 51(7): 1917-1928. (in Chinese)

[8] GUO J, BAO W, WANG J K, et al. A comprehensive evaluation framework for deep model robustness[J]. Pattern Recognition, 2023, 137: 109308.

[9] 王科迪, 易平. 人工智能对抗环境下的模型鲁棒性研究综述[J]. 信息安全学报, 2020, 5(3): 13-22.
WANG K D, YI P. A survey on model robustness under adversarial example[J]. Journal of Cyber Security, 2020, 5(3): 13-22. (in Chinese)

[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 (CVPR). Piscataway: IEEE, 2016: 2574-2582.

[11] WENG T W, ZHANG H, CHEN P Y, et al. Evaluating the robustness of neural networks: An extreme value theory approach[EB/OL]. (2018-01-31) [2024-12-01]. https://arxiv.org/abs/1801.10578v1.

[12] TIAN J Y, ZHOU J T, LI Y M, et al. Detecting adversarial examples from sensitivity inconsistency of spatial-transform domain[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2021, 35(11): 9877-9885.

[13] 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.

[14] MARCHETTI M, HO E S L. Improving Deep Learning Model Robustness Against Adversarial Attack by Increasing the Network Capacity[M]//Advances in Cybersecurity, Cybercrimes, and Smart Emerging Technologies. Cham: Springer International Publishing, 2023: 85-96.

[15] LING X, JI S L, ZOU J X, et al. DEEPSEC: A uniform platform for security analysis of deep learning model[C]// 2019 IEEE Symposium on Security and Privacy (SP). Piscataway: IEEE, 2019: 673-690.

[16] 李自拓, 孙建彬, 杨克巍, 等. 面向图像分类的对抗鲁棒性评估综述[J]. 计算机研究与发展, 2022, 59(10): 2164-2189.
LI Z T, SUN J B, YANG K W, et al. A review of adversarial robustness evaluation for image classification[J]. Journal of Computer Research and Development, 2022, 59(10): 2164-2189. (in Chinese)

[17] BASTANI O, IOANNOU Y, LAMPROPOULOS L, et al. Measuring neural net robustness with constraints[C]// Advances in neural information processing systems. Barcelona: NeurIPS, 2016: 29-37.

[18] ZHANG H, WENG T W, CHEN P Y, et al. Efficient neural network robustness certification with general activation functions[C]//Advances in neural information processing systems. Montréal: NeurIPS, 2018: 4939-4948.

[19] ZHANG C Z, LIU A S, LIU X L, et al. Interpreting and improving adversarial robustness of deep neural networks with neuron sensitivity[J]. IEEE Transactions on Image Processing, 2020, 30: 1291-1304.

[20] LIU A S, LIU X L, YU H, et al. Training robust deep neural networks via adversarial noise propagation[J]. IEEE Transactions on Image Processing, 2021, 30: 5769-5781.

[21] JIN H B, CHEN J Y, ZHENG H B, et al. ROBY: Evaluating the adversarial robustness of a deep model by its decision boundaries[J]. Information Sciences, 2022, 587: 97-122.

[22] KARIMI H, DERR T, TANG J L. Characterizing the decision boundary of deep neural networks[EB/OL]. (2020-01-03)[2024-12-01]. https://arxiv.org/abs/1912.11460v3.

[23] FAWZI A, MOOSAVI-DEZFOOLI S M, FROSSARD P. The robustness of deep networks: A geometrical perspective[J]. IEEE Signal Processing Magazine, 2017, 34(6): 50-62.

[24] HE W, LI B, SONG D. Decision boundary analysis of adversarial examples[C]//International Conference on Learning Representations. Vancouver: ICLR, 2018: 1-15.

[25] YU F X, QIN Z W, LIU C C, et al. Interpreting and evaluating neural network robustness[C]//Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence. Macao: IJCAI, 2019: 4199-4205.

[26] LEI S Y, HE F X, YUAN Y C, et al. Understanding deep learning via decision boundary[J]. IEEE Transactions on Neural Networks and Learning Systems, 2025, 36(1): 1533-1544.

[27] YAN J, YIN H L, ZHAO Z M, et al. Enhance adversarial robustness via geodesic distance[J]. IEEE Transactions on Artificial Intelligence, 2024, 5(8): 4202-4216.

[28] CHEN C, ZHANG J F, XU X L, et al. Decision boundary-aware data augmentation for adversarial training[J]. IEEE Transactions on Dependable and Secure Computing, 2023, 20(3): 1882-1894.

[29] KANBAK C, MOOSAVI-DEZFOOLI S M, FROSSARD P. Geometric robustness of deep networks: Analysis and improvement[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2018: 4441-4449.

[30] GOODFELLOW I J, SHLENS J, SZEGEDY C, et al. Explaining and harnessing adversarial examples[EB/OL]. (2015-05-20) [2024-12-01]. https://arxiv. org/abs/1412.6572v3.

[31] KURAKIN A, GOODFELLOW I, BENGIO S. Adversarial examples in the physical world[EB/OL]. (2017-02-11)[2024-12-01]. https://arxiv.org/abs/1607.02533v4.

[32] MADRY A, MAKELOV A, SCHMIDT L, et al. Towards deep learning models resistant to adversarial attacks[C]// 6th International Conference on Learning Representations. Vancouver: ICLR, 2018: 1-23.

作者简介

吴　涛　男,1987年出生于甘肃省庆阳市.现为重庆邮电大学网络空间安全与信息法学院教授、博士生导师.主要研究方向为人工智能安全、图数据挖掘、图机器学习、图模型安全等.中国电子学会会员编号:E190036235M.
E-mail: wutao@cqupt.edu.cn

汪俊杰　男,2001年出生于安徽省芜湖市.现为重庆邮电大学计算机科学与技术学院硕士研究生.主要研究方向为深度神经网络、模型对抗鲁棒性等.
E-mail: 2835094619@qq.com

曹新汶　男,1999年出生于重庆市.现为重庆邮电大学网络空间安全与信息法学院硕士研究生.主要研究方向为图神经网络、图模型安全、图模型对抗鲁棒性等.
E-mail: 398291796@qq.com

王　练　女,1976年出生于贵州省遵义市.现为重庆邮电大学网络空间安全与信息法学院教授、博士生导师.主要研究方向为软件安全、人工智能安全等.
E-mail: wanglian@cqupt.edu.cn

先兴平　女,1984年出生于四川省泸州市.现为重庆邮电大学网络空间安全与信息法学院副教授、硕士生导师.主要研究方向为图数据挖掘、数据隐私保护、智能算法安全等.
E-mail: xianxp@cqupt.edu.cn

张睿康　男,1999年出生于山西省太原市.现为重庆邮电大学计算机科学与技术学院硕士研究生.主要研究方向为深度神经网络、模型对抗鲁棒性等.
E-mail: 2362029949@qq.com