基于决策边界分析的深度神经网络鲁棒性评估与优先次序验证

林韧昊[1)] 周清雷[1)] 扈天卿[1)] 王一丰[2)]

[1)](郑州大学计算机与人工智能学院 郑州 450001)

[2)](信息工程大学密码工程学院 郑州 450001)

摘 要 随着深度学习技术在现实世界的广泛应用,人们对基于深度神经网络的系统安全性提出了更高要求.鲁棒性是神经网络的重要安全性质,对网络鲁棒性的量化分析和验证是深度学习模型安全性研究的关键问题.针对神经网络验证技术中难以解决的效率问题,提出了一种新颖的优先次序优化方法.结合局部鲁棒性的规约方式,在一组待验证输入内选择具有更高验证需求的不稳定点代替常规的逐点验证模式.根据对鲁棒性问题与决策边界距离的关联性分析,提出了一种基于网络输出单元值大小的鲁棒性评估方法作为优先验证的输入点选择依据.在此基础上将其扩展为输入的预分析模块与验证工具集成,进而设计了基于优先次序的验证框架.在常用的验证基准上进行了实验,结果表明,该方法的决策边界分析理论与突变测试结果一致,鲁棒性评估中选择不安全样本的平均准确率高于90%,通过减少安全样本的验证开销使验证效率提高了148.6%~432.6%.

关键词 深度神经网络;鲁棒性验证;优先次序模式;决策边界;鲁棒性度量指标

中图法分类号 TP311 **DOI号** 10.11897/SP.J.1016.2024.00862

Robustness Evaluation and Prioritization Verification for Deep Neural Networks via Decision Boundary Analysis

LIN Ren-Hao[1)] ZHOU Qing-Lei[1)] HU Tian-Qing[1)] WANG Yi-Feng[2)]

[1)](*School of Computer and Artificial Intelligence, Zhengzhou University, Zhengzhou* 450001)

[2)](*School of Cryptogram Engineering, Information Engineering University, Zhengzhou* 450001)

Abstract With the wide application of deep learning in the real world, people put forward higher requirements for the security of the systems based on deep neural networks. Robustness is an important safety property of neural networks, which is reflected in the vulnerability of models to adversarial perturbations. The quantitative analysis of network robustness is a key issue in the security research of deep learning models. Formal verification is an important technique to ensure the reliability of the models, using mathematical methods to construct rigorous encodings for models. Since neural networks have non-linear and large-scale structures, the existing verification technologies have intractable efficiency deficiencies. In view of this, a novel prioritization optimization method is proposed, which reduces verification time by introducing a pre-analysis process for inputs during verification to reduce the scale of the tasks. Specifically, combined with the limitations of local robustness specification, unstable points are defined as the inputs with higher verification requirements within a set of inputs to be verified, in instead of the conventional point-by-point verification mode. The proposed optimization method does not break the balance be-

收稿日期:2023-01-19;在线发布日期:2023-12-20.本课题得到国家重点研发计划项目资助.**林韧昊**,博士研究生,主要研究领域为人工智能安全、信息安全.E-mail:lrh417@gs.zzu.edu.cn.**周清雷(通信作者)**,博士,教授,中国计算机学会(CCF)杰出会员,主要研究领域为形式化方法、人工智能算法.E-mail:ieqlzhou@zzu.edu.cn.**扈天卿**,博士研究生,主要研究领域为图像处理、机器学习.**王一丰**,博士研究生,主要研究领域为网络安全、机器学习.

tween verification accuracy and efficiency. In order to accurately select unstable points that are prone to unsafe, the causes of model robustness problems are analyzed in detail from the perspective of decision boundaries, involving the generalization ability and overfitting issues. Points that are closer to the decision boundary are more likely to misclassify the neural network when perturbed. Then, according to the correlation analysis of the robustness problem and the distance of the decision boundary, a robustness evaluation method based on the value of the network output unit is proposed as the input point selection basis for priority verification. A lightweight robustness metric based on the output difference is defined to reflect the distance relationship between the input point and the decision boundary, and advanced extension forms are presented. Also, the theoretical basis for this metric is provided in terms of adversarial attack and defense, and mutation testing. On this basis, the input pre-analysis module is extended to integrate with the verification tools. Furthermore, a prioritization-based verification framework is designed, and the working principle and specific process of the framework are demonstrated. The integration ways and implications of proposed method in different types of verification tools are discussed from practical application of view. Extensive experiments on commonly used verification benchmarks demonstrate the rationality and effectiveness of the proposed method. The accuracy of the metric is proved by comparing the consistency of the output results of different input points divided based on the robustness evaluation method in strict formal verification tools. The selected unstable points are sequentially used as representatives of input points that need to be heavily considered in the verification, and increasing efficiency by ignoring points that are probabilistically safe during the execution of the tools. The results show that the decision boundary analysis theory is consistent with the results of mutation testing, the average accuracy of selecting unsafe samples in robustness evaluation is higher than 90%, and the verification time is reduced by 148.6%～432.6% by declining the verification costs of safe samples.

Keywords　deep neural network; robustness verification; prioritization; decision boundary; robustness metrics

1　引　言

近年来,深度神经网络(Deep Neural Networks,DNN)[1]凭借其强大的表达能力,在计算机视觉[2]、语音识别[3]、自然语言处理[4]等领域得到了广泛的应用.然而,尽管神经网络能够有效解决诸多复杂问题,但是它们在安全关键型系统(Safety-Critical System)[5-7]中的应用仍然受到限制.这是因为DNN模型存在严重的鲁棒性问题,即非常容易受到对抗攻击的威胁[8].具体来说,在神经网络的输入中施加人眼难以察觉的扰动就会使模型以一个很高的置信度产生截然不同的误分类结果,这样的输入称为对抗样本(Adversarial Examples, AE)[9].在具有较高安全性需求的系统中,这种隐蔽性极强的攻击可能会引发难以估量的损失,使得DNN仅限于满足最低安全级别的系统.

随着深度学习应用的日益广泛,鲁棒性问题带来的安全威胁日益突出.为此有必要为神经网络模型提供严格的可信性保证.形式化验证是模型安全性分析的可靠手段[10],但是神经网络本质上是一个难以解释的黑匣子,使得传统的形式化技术无法直接有效地应用于DNN模型的证明.目前面向神经网络的验证技术主要可分为完备(Complete)和不完备(Incomplete)的验证[11].完备的验证通常以基于可满足性模理论(Satisfiability Modulo Theories, SMT)[12-13]、混合整数线性规划(Mixed Integer Linear Programming, MILP)[14-15]和符号区间传播[16]等方法为代表.此类方法通常依赖于一些约束求解工具,通过对网络进行精确的编码,能够确定地阐述某个安全性质是否成立,但具有昂贵的计算开销.不完备的验证以抽象解释[17-18]、凸松弛[19]、线性近似[20]、区间边界传播[21]和Lipschitz常数[22]等方法为代表,通过引入过近似松弛,能够更加快速地验证

更大规模的神经网络,但有时只能给出一个有界估计,无法直接给出性质是否成立的保证.两类方法都是 DNN 可靠性保证的重要技术,具体取决于实际应用领域的需求.

目前 DNN 的鲁棒性验证主要存在效率及可扩展性问题.即使对于具有良好可扩展性的不完备验证方法,也难以在精度和效率上取得良好的平衡,因为较高验证精度伴随的计算复杂度问题自然无法避免.若使用过多的近似追求验证效率,则会丧失验证本身的意义.为了使形式化方法更好地应用于 DNN 安全性证明,不仅需要研究高效的验证算法,还要通过制定针对性的辅助优化策略提高验证性能.目前广泛使用的局部鲁棒性的局限性在于一次只能考虑神经网络在一个给定样本邻域内的稳定性,这使得验证工具主要采用逐点验证模式(Pointwise),并且需要对大量的输入点进行随机测试.因此,这种验证模式容易受到 DNN 验证效率的限制且不具备说服力.

本文围绕一个新颖的优化方向—优先次序(Prioritization)[23]展开研究.针对目前验证工具的性能问题,若能够从一组待验证输入内选择最有可能违反模型性质的鲁棒邻域中心点进行优先验证,就能节省常规逐点的批量验证所带来的大量计算开销,尤其在具有更高精度的验证工具中更有利于效率上的提升.具体来说,这需要在验证前对神经网络上不同输入点的鲁棒性进行评估作为目标优先级的选择依据,通过在大规模输入集内对各输入点不满足安全性质的概率进行排序来定位其中最值得被优先关注的不稳定点,从而减少待验证的任务规模并提高验证效率.主要贡献如下:

(1)针对局部鲁棒性形式规约的局限性提出了一种基于优先次序的优化方法,通过选择具有较高验证需求的输入点进行优先验证来代替传统的逐点验证模式,并提高局部鲁棒性验证的代表性.

(2)基于决策边界分析提出了一种基于输出单元值的鲁棒性度量指标来选择不稳定点作为优先次序模式的实现基础,并分别从对抗攻防和突变测试的角度提供了理论依据.

(3)将鲁棒性评估实现为易集成的输入预分析模块,设计了一种基于优先次序的验证优化框架,并对不同类型验证工具的集成进行了讨论.

(4)在常用的验证基准中进行了实验.分别从决策边界分析合理性、鲁棒性评估准确性、验证效率提升三个方面证明本文方法的有效性.

2 背景和相关工作

2.1 深度神经网络结构

一个神经网络模型 $f:\mathbb{R}^{n_0}\rightarrow\mathbb{R}^{n_k}$ 可以被看作是两个高维向量空间的映射,由多个层(Layer)组成,层与层之间由相互连接的神经元节点组成.DNN 通过权重矩阵 $\boldsymbol{W}$ 和偏执向量 $\boldsymbol{b}$ 对输入 x 进行仿射变换,并通过激活函数 $\sigma:\mathbb{R}^{n_l}\rightarrow\mathbb{R}^{n_{l+1}}$ 增强神经网络的表达能力.一种在验证中受欢迎的激活函数是 *ReLU*:$\sigma(x)=\max(0,x)$,因为其分段线性特性是相对易于验证的.*DNN* 模型可以由以下递归方程描述:$x^0=x, x^l=\sigma^l(\boldsymbol{W}^l x^{l-1}+\boldsymbol{b}^l), l=1,\cdots,k, f(x)=x^k$,其中 $x^0\in\mathbb{R}^{n_0}$ 为网络的输入,$x^l\in\mathbb{R}^{n_l}$ 为第 l 层的输入,$f(x)\in\mathbb{R}^{n_k}$ 为网络的输出.分类器把得分最高的类作为预测标签 $C(x)=\mathrm{argmax}_{c_i} f_{c_i}(x)$,其中 $f_{c_i}(x)$ 是输出向量 $f(x)$ 中标签 c_i 对应的第 i 类概率分量,$i\in\{1,\cdots,n\}$.

区别于传统的软件架构,对 *DNN* 的验证主要面临非线性和大规模两个挑战.首先,激活函数虽然赋予了 *DNN* 强大的拟合能力,但同时也引入了非线性因素导致神经网络的验证成为了一个大规模非凸问题,使得直接计算其输出范围非常困难.模型参数由学习过程中的优化算法决定,缺乏可解释性和形式化的逻辑结构,给模型的编码也带来了困难.其次,随着神经网络的应用场景越来越复杂,表现优秀的模型通常有较大的规模.这样的神经网络结构更为复杂,包含大量的隐藏节点.以上两种特性直接引发了 *DNN* 验证的效率和可扩展性问题,使得验证一个简单的分段线性网络已经是一个 *NP* 完全(*NP-Complete*)问题[12],为此有必要通过针对性的优化方法提高验证效率.

2.2 鲁棒性验证

形式化验证基于数学推导为模型提供严格的安全性证明.通常 DNN 模型的可靠性验证需要对神经网络 f 和安全性质 P 进行建模和规约.一个验证查询可以表示为 $V:=<f,P>$ 形式,其中网络 f 被编码为数学公式,性质 P 包含输入和输出约束.从规约角度,验证问题可以归结为在多维有界的输入邻域内 $\eta=\{l_d\leqslant x_d\leqslant u_d \mid l_d,u_d\in\mathbb{R}\}, d\in\{1,\cdots,m\}$,寻找网络的输出上某些函数的全局极小值[24],例如真实标签 C 和其他目标标签 $c_i\neq C$ 预测之间的差异.

鲁棒性作为神经网络重要的安全性质,表示输入的微小变换不会导致模型决策出现较大偏差,主要基于对抗性扰动 δ 定义.大多数研究将其分为局部鲁棒性和全局鲁棒性.其中局部鲁棒性刻画了神经网络在一个正常样本 x 的邻域内的行为是一致的.输入的邻域 η 由 L_p 范数 $\|\cdot\|_p$ 定义,$p\in[1,+\infty]$,验证中最常用的是一个以 x 为中心 δ 为半径的 L_∞ 范数球 $B_\infty(x,\delta):=\{x'\mid\|x'-x\|_\infty\leqslant\delta\}$(凸多面体).以 Katz 等人[23]的定义为例,神经网络 f 在输入 x 处的局部鲁棒性可表示为:$\forall x'\in B_\infty.\ |f_{c_i}(x')-f_{c_i}(x)|<\varepsilon$,其中常数 $\delta>0$ 和 $\varepsilon>0$,或者根据网络的输出 $f_C(x')>f_{c_i}(x')$ 来判定.这里标签 $c_i\in L$ 对应的输出节点值 $f_{c_i}(x)$ 可称为网络中 x 被标记为 c_i 的置信度,L 为分类标签集合.全局鲁棒性能够刻画神经网络对整个输入域 D 的稳定性,即:$\forall x,\ \widetilde{x}\in D.\ \|x-\widetilde{x}\|\leqslant\delta$,$|f_{c_i}(x)-f_{c_i}(\widetilde{x})|<\varepsilon$,其中 x 和 $\widetilde{x}$ 为任意两个距离足够近的输入.

相较于全局鲁棒性,局部鲁棒性更易于验证,但其局限性在于只能刻画一个小范围的输入邻域.考虑到全局鲁棒性对于现阶段的验证工具过于复杂,会引发更严重的效率问题,使得目前的验证工具主要关注局部鲁棒性的验证.但是在以某个单一点为中心的局部验证中,随机选取样本点会导致这种验证模式难以具备代表性.全局输入空间可以看作是由大量局部邻域组成,如果能从中选择出所对应的更容易违规的原始输入点 x,就能在一定程度上实现局部代表全局的验证.

2.3　验证技术现状

深度神经网络的鲁棒性威胁严重限制了其在安全攸关系统中的应用.在使用 DNN 模型执行智能任务之前,必须保证其鲁棒性满足应用需求.在神经网络安全领域,形式化验证是证明网络行为正确性的一项重要技术.我们总结了一些现阶段典型的验证方法,如表 1 所示.

完备的验证需要对整个神经网络进行精确编码,原则上可以解决任何验证问题.虽然可以使用符号区间传播等技术优化网络输出上下界的计算过程,但仍受到验证算法计算复杂度的限制,一般只能支持分段线性激活函数、简单的图像数据集和规模较小的模型.不完备的验证为网络的编码引入近似处理,有较好的可扩展性,同时也会产生一定的误差,在保证精度的前提下效率仍受到限制.对于 DNN 验证的优化大多集中在验证算法层面,往往会造成精度和效率其一的损失,我们在之后基于验证模式特点展示了一种新的优化思路.

表 1　DNN 鲁棒性验证方法

工具	完备	模型	激活函数	关键技术
Planet[13]	√	FNN	R	SMT
Reluplex[12]	√	FNN	R	SMT
Marabou[25]	√	FNN,CNN	R	SMT
Tjeng etal.[14]	√	FNN, CNN	R	MILP
Neurify[16]	√	FNN	R	符号线性松弛
β-Crown[24]	○	FNN,CNN	R,S,T	边界传播
Verinet[11]	○	FNN, CNN	R,S,T	符号区间传播
DeepPoly[17]	×	FNN,CNN	R,S,T	抽象解释
k-poly[18]	×	FNN,CNN	R	抽象解释
FastLin[20]	×	FNN	R	线性近似
FastLip[20]	×	FNN	R	Lipschitz 常数

注:○表示同时具备完备和不完备模式;激活函数:R 表示 ReLU,S 表示 sigmoid,T 表示 tanh.

3　基于输出单元值的鲁棒性评估

本节基于决策边界(Decision Boundary, DB)分析提出了一种鲁棒性度量指标作为后续优先次序验证的实现基础.使用神经网络输出层各单元值的大小评估输入点与决策边界的距离,以度量模型对不同输入点的鲁棒性,并结合基于决策边界的 Deepfool 攻击以及突变测试(Mutation Testing, MT)技术等多个角度对所提方法进行了讨论.

3.1　决策边界分析

首先,我们研究了神经网络鲁棒性问题产生的原因.本文的鲁棒性评估主要依赖于决策边界问题分析.决策边界 $\mathcal{D}$ 作为分类的基础,是一种假设函数的属性,由训练好的模型参数决定,对任意一对不同类别 $\{c_i,c_{i^*}\}$ 有 $\mathcal{D}_{\{c_i,c_{i^*}\}}\triangleq\{x\mid f_{c_i}(x)=f_{c_{i^*}}(x)\}$.神经网络的鲁棒性通常指输入点在 L_p 范数度量的局部邻域内是否存在被错误分类的情况,而靠近 DB 的脆弱样本会优先不满足鲁棒性.以一个二分类问题为例,图 1 展示了一个简单模型的二维决策边界.从图中可以看出,距离 DB 越近的点,其 L_p 范数球 $B_p(x,\delta)$ 区域内更可能发生错误预测(图中 $p=\infty$).我们将此类输入点定义为不稳定点(Unstable Point, USP) $\bar{x}$,因为它们受到微扰时更加不稳定,即只需要很小的扰动就能使 USPs 跨越对应的分类边界超平面.此外,对于多分类模型同样可以基于多条 DBs 的分析来评估 DNN 对输入点的鲁棒性.

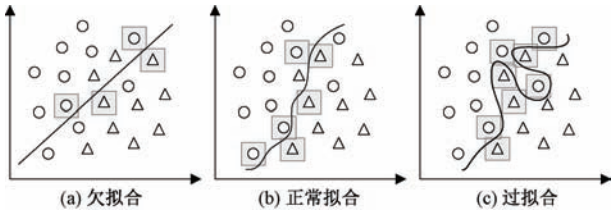

图 1 决策边界示意图

一般认为造成 DNN 模型的鲁棒性缺陷的一个主要原因就是过拟合(Overfitting)现象. 图 1 中从左到右分别对应了模型的欠拟合、正常拟合和过拟合现象. 一般在神经网络的训练阶段,为了保证训练精度,会不断拟合训练数据,导致决策边界收敛出现过拟合形状,进而存在更多位于边界附近的输入点. 而距离 DB 越近的点越容易受扰动的影响,同时导致网络对未观测过的样本也更容易出错. 过拟合问题能够反映 DB 分析理论的合理性.

3.2 鲁棒性度量指标

基于决策边界分析,本节提出了一种轻量级的鲁棒性评估方法. 核心依据是越靠近 DB 的点,在基于 L_p 范数度量的局部邻域 $B_p(x,\delta)$ 内越容易跨越 DB 输出错误分类结果. Feng 等人[26]从促进反例生成的角度提出了一种验证的改进方案,他们认为如果神经网络的最大输出值与第二大输出值相差不大的情况下更容易产生反例. 在基于 Lipschitz 常数的方法[27]中同样使用了鲁棒邻域中心点 x 输出中各个类的输出值来分析神经网络的行为,且网络函数本身及其输出都是 Lipschitz 连续的[28].

本文希望基于神经网络输出层单元值大小来评估输入点到不同决策边界之间的距离,从而实现对模型在不同输入点邻域 η 内的鲁棒性度量,并在 4.3 和 4.4 节中为二者的相关性提供了证明. 这里我们主要关注多分类神经网络,通常此类模型的输出层节点数量 $\{y_1,y_2,\cdots,y_n\}$ 对应输出标签数量 $\{c_1,c_2,\cdots,c_n\}$,利用 DB 超平面对数据点所属类别的区域进行划分. 我们认为 DNN 输出层中各个节点的输出值大小能够反映 x 与 DB 间的距离,若存在一个或多个输出分量值与真实标签对应的最大输出分量值更加接近的点,则认为该输入点 x 位于网络 f 的 DB 附近,那么 f 对其的鲁棒性也相对更弱.

神经网络 $f:\mathbb{X}\rightarrow\mathbb{Y}$ 是输入到输出的映射,其输出层包含 n 个节点,每个节点对应一个标签 c_i. 我们使用 $y_i=f_{c_i}(x),i\in\{1,\cdots,n\}$ 表示第 i 个输出单元值(softmax 之前)并按照其大小进行排序,例如 y_1 和 y_i 分别表示最大和第 i 大的输出单元值. 通常 y_i 越大,输出对应标签 c_i 的概率就越大,且 y_1 对应了 x 的真实标签 $C(x)$. 本节基于神经网络各输出单元的差值定义了度量指标 ς 来评估模型在不同中心点上的鲁棒性. 考虑到局部鲁棒性的定义方式,被施加微小扰动的输入更有可能跨越距离最近的 DB,因此我们优先关注间隔最小的情况,即 $\varsigma(x)=y_1-y_2$. $\varsigma(x)$ 越小则 f 在 x 附近有错误行为的概率越高,我们设定了阈值 H_{us} 来选择满足条件的次大输出值:$\varsigma(x)<H_{us}$,并将对应输入判定为不稳定点 $\bar{x}$,作为需要被优先验证的初始点. DNN 的鲁棒性和输出差值会受到模型结构、训练数据和参数等因素的影响,在实际模型间可能存在较大差异. 在不同网络中需预先对多组输入执行前向传播得到它们的输出差值范围并指导 H_{us} 的设置和调整,该操作的时间开销几乎可以忽略不计. 从评估的角度,可以估计输入点在决策面中的分布情况,从优先次序验证的角度基于 $\varsigma(x)$ 升序排序即可,而阈值 H_{us} 可用于灵活地控制不稳定点的数量和比例,以及省略稳定的输入. 另外,在鲁棒性验证中只关注 y_1 对应正确分类标签的样本.

多分类模型有多条决策边界,如果存在多个超过阈值 H_{us} 的输出值 y_i,可以认为该点周围存在多条距离较近的 DBs. 从对抗攻击的角度来看,根据其是否需要输出指定的误分类标签可分为目标攻击和非目标攻击. 对于目标攻击的鲁棒性有 $\varsigma(x)=y_1-y_i$,用于分析 x 受到对抗性扰动时输出指定误分类 $i\neq1$ 的概率. 通过输入点到不同决策边界之间的距离估计选择合适的攻击样本,根据攻击方向对 $\bar{x}$ 只需要更小的扰动就能实现更高的目标攻击成功率. 非目标攻击只需要关注输入点最接近的 DB,考虑 $f(x)$ 中评分次高的维度 y_2 即可. 值得注意的是,$\varsigma(x)$ 主要从高效的输入预选择角度出发,大致评估模型在大量输入约束下不满足局部鲁棒性的概率,在精确度方面可能会有一些误差. 而真实的 DB 距离(Margin)评估还需要考虑模型自身的仿射变换参数,可以被表示为神经网络 logit 输出的梯度形式,即 $\lambda_i=\|\nabla_x f_{c_1}(x)-\nabla_x f_{c_i}(x)\|_q^{-1}$,来解释输出层每个神经元对决策方向的影响[29-31],其中 L_q 是 L_p 的对偶范数,$q\triangleq\frac{p}{p-1}$. 此时可以使用 $\varsigma^*(x)=\min\{\lambda_i\varsigma(x)\}$ 获得更精确的距离度量和分析扰动范围的对应关系. 而实验中表明 $\varsigma(x)$ 已经足够用于不稳定点的选择,这可能是因为这种梯度信息 λ_i 通常不会对选择结果造成明显差异. 在比较距

离同一边界的同类点,以及观测网络行为在某一样本上受扰动的影响时可以省略这种因素.

基于 δ -局部鲁棒性的定义以及 $\varsigma(x)$ 中 y_i 与决策边界距离的相关性这一结论,同样可以尝试在 L_p 范数球中(如 B_∞)为 x_d 采样多组值执行前向传播观察 x' 各输出值 $y'_i=f_{c_i}(x')$ 的变化,对不同原始输入的鲁棒性做进一步分析. 首先这需要沿着合适的决策边界方向寻找赋值,即 $\varsigma(x)$ 下降情况,或者依赖于一些对抗攻击. 如果在给定 δ 下 $C(x)\neq C(x')$,之后可以定义 $\varsigma'(x')=\max_{i\neq 1}(y'_i-y'_1)$,其中 y'_i 和 y'_1 分别为 x' 改变后的预测标签和 x 的原始标签对应的输出层节点值. 这反映了扰动后 x' 跨越正确决策边界的幅度, $\varsigma'(x')$ 越大表明 f 在 x 处受扰动的影响越大. 然而在微扰下,上述情况一般属于小概率事件,即使不能满足误判条件,也可以继续使用 $\varsigma(x')$ 观测 $\Delta(\varsigma(x),\varsigma(x'))$ 来具体分析输出得分的变化,并且从目标攻击角度可以考虑多个原来评分靠前的良性维度,如排名前三的 $\{y_i\}_{i=1}^3$. 另外,基于不同 δ 下的度量差异有助于估计 B_p 内更有威胁的扰动方向以及真实可接受的最大扰动范围(由 y'_1 和 y'_i 更接近的情况指导),但不如形式化方法得到的精确结果. 虽然 x' 考虑了 δ 值的大小,但会增加额外的度量成本.

与正式的验证框架不同,本文的评估主要用于在不同输入下估计给定模型的违规概率,优势是有着极低的计算成本(一次前向传播即可). 在局部搜索空间中,理论上可以使用 $\varsigma(x)=\min_{i\neq 1}(y_1-y_i)$ 作为优化问题寻找区域内的最坏情况,基于约束 $P_{out}:=f_C(x')>f_{c_i}(x')$ 来检测性质成立与否,并通过最小化 $\varsigma(x)$ 寻找反例. 此外,若为了实现一种量化的分析,可以通过计算 f 在 L_p 范数球 $B_p(x,\delta)$ 上的局部 Lipschitz 常数 L_B [27],得到最小扰动的下界 $r_{min}^L=\min_{i\neq 1}\dfrac{\varsigma(x)}{L_B}$,以保证分类器对任意扰动 $\|r\|_p\leqslant r_{min}^L$ 是鲁棒的,其中 $r\in B_p(x,\delta)$. r_{min}^L 定义了不会导致网络行为异常时 x 可接受的最大扰动,这也反映出 $\varsigma(x)$ 和鲁棒性的紧密联系. 但是该方法的精度往往较差,因此本文倾向于从中心点优先级排序的角度出发,通过更轻量级的评估用于集成更高精度的验证工具.

3.3　对抗攻防分析

本节基于经典的 Deepfool[30] 对抗攻击方法说明分类器输出值和决策边界之间的关系. Deepfool 通过沿着 DB 方向迭代施加扰动 r 生成对抗样本 $x'=x+r$, $\|r\|_p\leqslant\delta$,使得 $C(x+r)\neq C(x)$. 该方法根据点到直线的距离:

$$dis=\left|\frac{Ax_1+Bx_2+C}{\sqrt{A^2+B^2}}\right|,$$

实现对二维输入 (x_1,x_2) 的位移,进而推导出多分类器中点 x 到最近边界超平面的距离:

$$\hat{l}(x)=\operatorname{argmin}_{c_i\neq C(x)}\frac{|f_{c_i}(x)-f_{C(x)}(x)|}{\|w_{c_i}-w_{C(x)}\|_2},$$

这里 c_i 表示改变后的输出类别, $f_{c_i}(x)$ 为误分类标签的概率分量, $C(x)$ 表示样本的原所属类别, $f_{C(x)}(x)$ 为 x 的真实分类标签对应的概率分量.

训练后的神经网络中权重参数 w 是固定的,由 $\hat{l}(x)$ 的计算公式可得,对应误分类标签和真实标签的输出节点值的差值 $|f_{c_i}(x)-f_{C(x)}(x)|$ 越大,说明输入 x 到误分类决策边界的距离越远. 由于 x 在模型 f 中的预测类别为 $C(x)$,所以 $f_{C(x)}(x)>f_{c_i}(x)$,可得 $f_{C(x)}(x)=y_1$ 越大, $f_{c_i}(x)=y_i$ 越小,使网络对 x 误分类所需的扰动 r 就越大,此时 f 在 x 上的鲁棒性也就越强,并且 y_2 所对应的是距离 x 最近的 DB. 由此可以得到 $\varsigma(x)$ 正比于 $\hat{l}(x)$,为输出层各节点的输出值大小可以度量输入点与决策边界之间的距离关系提供了直接的理论保证.

从防御角度来说,对抗训练[32]可以表示如下:

$$\min_\theta\mathbb{E}_{(x,c)\sim\mathcal{D}}\left[\max_{\|r\|\leqslant\delta}L(x+r,c;\theta)\right],$$

其中 L 为损失函数, θ 为对抗样本期望风险最小的模型参数.

文献[33]使用 PGD 对抗训练进行测试,虽然训练后模型的分类准确率有所下降,但是泛化能力有明显提升. 模型参数的变化会对模型的 DB 和输出造成影响,一般由模型的结构设计、训练数据和算法决定. 例如将对抗样本注入训练集对参数正则化来防止过拟合,使得数据点远离决策边界[34],此时我们观察到随着 DB 距离的增加对应指标 $\varsigma(x)$ 也有相应的增加,间接反映了它们之间的关系.

3.4　突变测试分析

本节通过突变测试[35]进一步说明决策边界、鲁棒性问题以及输出节点值之间的关系. MT 是一项通过良好定义的突变操作发现传统软件缺陷的测试技术. Ma 等人[36]提出了面向深度学习模型的 MT,通过模型级变异算子(Mutation Operator)不经过训练阶段直接将错误注入 DNN 来评价测试数据的质量. 我们在另一项工作[37]中提出了基于 MT 的决策边界距离与鲁棒性评估方法,使用模型级突变算

子改变原始模型的内部结构生成大量突变模型(Mutation Model, MM),从而对神经网络的原始DB进行微调.如果原始模型 f 和突变模型 $\widetilde{f}$ 对同一个输入 x 的决策产生偏差,那么网络的DB显然得到了变化.我们以一个适当的突变率 R_M 生成足够多的MMs有助于分析输入点到决策边界的距离关系.

我们可以通过突变测试改变待验证模型的决策边界以证明指标 ζ 的合理性.具体来说,可以使用高斯模糊(Gaussian Fuzzing,GF)和神经元激活翻转(Neuron Activation Inverse,NAI)等变异算子获得大量的突变决策边界.如图2所示,若某个输入点能够多次跨越同一类的突变决策边界,这表明该样本与原始模型中所对应的DB距离较近.根据突变测试的结果,如果基于 $\zeta(x)$ 选择的位于DB附近的输入点 $\bar{x}$,在生成后的突变模型 $\widetilde{f}(x)$ 中能够多次输出该边界对应的误分类标签 c_i ,则表明 $\zeta(x)$ 与MT中DB距离度量结果的一致性.以MNIST[38]数据集训练的DNN为例,该模型有10个输出单元,分别对应数字0－9.假设一张人眼识别为3的图像,在输出层中的第二大和第三大输出值对应的标签为5和2.此时我们希望该图像距离标签5和2对应的DB更近,即在一定 R_M 下生成的MMs输出最多的错误类别为5和2,并在实验中证明了这一点.

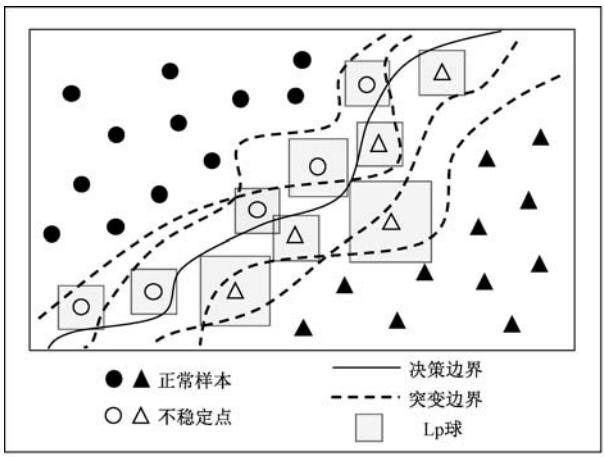

图2 MT评估输入点与DB的距离

4 优先次序验证设计

本节基于优先次序思想提出了一种新的验证框架,将鲁棒性评估作为一个可集成的验证工具辅助优化模块,展示了该验证模式的整体工作流,并对其实际验证中的集成进行了讨论.

4.1 优先次序验证

本节介绍了优先次序思想对于DNN模型验证的意义.现有的验证主要以神经网络自身作为验证模型,由于DNN的非线性及大规模特点,使得验证效率问题难以得到根本上的解决.在神经网络的鲁棒性规约方面主要基于输入和输出约束 $P_{in}\Rightarrow P_{out}$ 的形式,因此执行一次验证得到的是模型对某个输入点的性质是否成立.局部鲁棒性的验证就是通过对多个输入点可接受的扰动半径进行检验,从而实现对模型整体可靠性的评估.但众所周知神经网络的内部结构主要取决于人为经验设定,其内部参数的生成取决于网络训练过程中的优化算法,使得模型对于不同输入点的鲁棒性强弱有所差别.

对于一个训练完备的DNN模型,待验证的输入不仅要包含训练样本,还应包含未曾观测过的样本.为了在实际验证中对模型的鲁棒性进行更全面的分析,往往需要处理数量庞大的输入样本,但这对于受到计算复杂度问题制约的形式化方法非常困难.现阶段的验证工具大多都采用随机选取测试点进行逐点验证,显然难以具备说服力,并且局部鲁棒性只能刻画神经网络在一个输入点邻域内的稳定性,使得这种验证模式中存在极大的偶然性.在一些具有较高精度的完备验证工具中,验证简单网络在一个输入点上的鲁棒性就要花费数百甚至数千秒的时间,给形式化方法在实际规模的DNN模型中的应用带来了巨大的阻碍.

针对上述问题,本文提出了一种基于优先次序的DNN鲁棒性验证优化方案.其主要思想为通过在一组输入内选择最具有验证价值的点作为代表,以代替传统的逐点验证模式.我们认为优先验证以相对不安全的输入点为中心的鲁棒邻域会更有意义.虽然人们期望神经网络能够满足鲁棒性质,但由于模型训练中的泛化能力(Generalization)不足等原因使这种先天缺陷难以避免.从防御角度来说,对抗训练[32]是提高DNN鲁棒性最有效的手段之一,然而使模型对不同样本产生误判行为所需的扰动幅度不同,选择容易受到攻击威胁的样本能生成高质量的训练数据,不会造成大的模型准确性缺陷[6].因此在DNN的鲁棒性研究中,更有可能不满足安全性质的点有更高的验证价值,更有助于分析模型中的鲁棒性问题和指导可靠的模型训练[16,39].我们可以将其归为神经网络对应的庞大输入数据中一类特殊的脆弱样本,由3.2节的鲁棒性评估方法指导选择.

在验证算法层面,如何平衡其精度和效率是一个难题.这是因为验证中涉及到大量非线性计算,随着模型越来越复杂,在保证验证精度前提下使得其效率难以被兼顾.对于多种高精度验证工具,处理大批量输入样本有着极大的计算开销,而从中选择具有代表性的点并舍弃验证需求相对较低的点可以有效减少验证中的任务数量,这样可以在有限的计算资源下关注更多的脆弱样本.另外,完备的验证能够在输入不满足性质时产生一个反例,从而直接用于对抗性训练提高神经网络的抗扰动能力.所以优先次序思想对于提高验证结果说服力以及验证效率方面都有重要作用和价值,尤其对于有着大量验证查询和较高精度需求的验证任务.

4.2　验证框架

为了更好地对神经网络模型的可靠性进行分析,初始输入点的选择非常重要.无论对于模型的训练或推理阶段往往都需要依赖大量的输入数据作为支撑,因此待验证输入集一般具有很大的规模.在面对大批量的验证查询时,随机选取测试点的验证模式显然难以胜任.考虑到验证工具的性能瓶颈,对大量输入的逐点验证是难以实现的.

通常验证在给定扰动下容易出错的点更有助于分析模型中存在的鲁棒性威胁,从输入选择的角度在有限的计算资源中优先选取更有可能违背 DNN 鲁棒性的点进行验证.由于指标 $\varsigma(x)$ 能够反映各输入点与决策边界之间的距离,我们根据距离误分类边界越近的点越容易不满足鲁棒性质作为不稳定点的选择依据,并将优化策略作为一个预分析模块与验证工具进行集成提出了一个基于优先次序的验证框架,如图 3 所示.

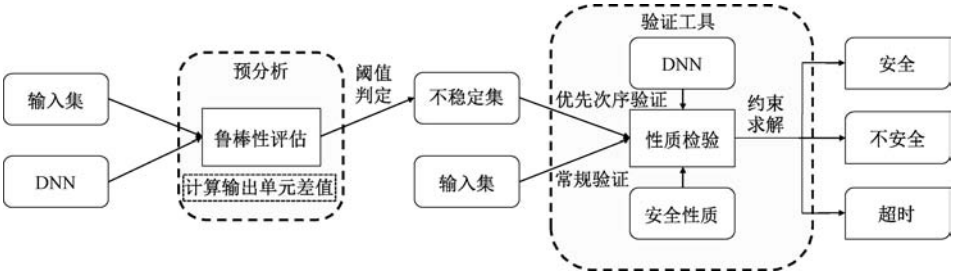

图 3　集成后的验证框架

常规的验证需要对模型及其性质进行编码,通过它们之间的可满足关系进行判定.我们为传统的验证框架引入了一个输入的预分析过程,用于从待验证的输入集合 X 内选择不稳定点 $\bar{x}$ 作为代表.将每个输入的 $\varsigma(x)$ 值大小作为验证优先级的排序指标,以此来评估输入点与神经网络决策边界的关系,距离越近对扰动越敏感.对于 $\varsigma(x)$ 值较大的点我们认为其在受到扰动时更加稳定.随后可以根据验证结果对模型的鲁棒性进行评价,若其在绝大多数不稳定点 $\bar{x}$ 处对性质 P 是可满足的(Satisfiable, SAT),可以大致判定该网络对该输入集是 δ −局部鲁棒的.这里我们将判定条件设置为 $N_{SAT}>\alpha N_{us}$,其中 $\alpha\in[0,1]$ 为比例系数,N_{us} 表示不稳定点集合 $\bar{X}$ 中 $\bar{x}$ 的数量,N_{SAT} 表示满足性质的 $\bar{x}$ 数量.若大部分 $\bar{x}$ 是不满足的(Unsatisfiable, UNSAT),我们可以通过放宽 H_{us} 的设置进一步选择更多输入样本进行验证以生成更多可能的反例,同时也可以判断该模型在给定 δ 下的鲁棒性较弱.值得注意的是,很多验证工具(如 Reluplex[12]、Neurify[16] 等)对性质采用否定编码的形式,在违反安全属性时的输出为 SAT.为了描述方便,本文统一将不满足结果记为 UNSAT.

本文优化的核心思想是在验证前对不同输入点的鲁棒性进行预分析.在作为一个独立的鲁棒性评估方法进行使用时,同样可以根据满足 USP 判定条件的输入数量实现对模型整体鲁棒性的分析.与形式化方法不同,神经网络的输出值大小是由其内部结构及参数决定的,基于输出值的评估虽然无法对网络的具体性质进行确定性的量化推理,但基于违反 δ −局部鲁棒性质的输入比例估计,可以有效用于选择某个特定网络上更易受对抗扰动影响的样本.由于鲁棒性关注的是训练完备模型的推理阶段,所以 USPs 的数量占比只取决于鲁棒性度量结果.但随着 $B_{\infty}(x,\delta)$ 邻域内半径 δ 设置的改变,验证工具对于同一个样本点的验证结果也会发生变化.因此在集成时应该使 δ 在合适的范围内,在确保扰动隐蔽性的同时避免验证和选点丧失意义.如果 USPs 在 δ 较小时能够违反安全性质,则可以证明指标 ς 的有效性.

4.3　验证算法

我们基于优先次序验证框架展示了优化后的整

体验证工作流，如算法 1 所示.

算法 1. 优先次序验证模式.

输入：神经网络 f，输入集 X，阈值 H_{us}，验证查询 V.

输出：神经网络 f 在性质约束 P 下是否鲁棒.

阶段 1 输入预分析.

1. Initialization
2. FOR $x_j \in X$ DO
3. $y_i \leftarrow f_{c_i}(x_j)$
4. Update i in order of y_i from high to low
5. IF $\varsigma(x_j) < H_{us}$ THEN
6. Add $\bar{x}_j$ to $\bar{X}$；
7. ELSE $j \leftarrow j+1$
8. Update $x_j \in \bar{X}$ in order of $\varsigma(x_j)$ from low to high
9. RETURN $\bar{X}$

阶段 2 验证.

10. WHILE true DO
11. Submit $\bar{x}_j \in \bar{X}$ to verification tool in order
12. Solve $V(f,P)$
13. IF $f_C(\bar{x}'_j) > f_{c_i}(\bar{x}'_j)$ THEN
14. $N_{SAT} \leftarrow N_{SAT}+1$
15. IF $N_{SAT} > \alpha N_{us}$ THEN
16. RETURN f is safe under P
17. ELSE
18. RETURN f is unsafe under P
19. Reset $H_{us} \leftarrow H_{us}^{+}$
20. END

阶段 1 展示了预分析过程. 首先在模型 f 中对待验证输入 $x_j \in X$ 执行前向传播计算输出层各单元值 y_i，同时对 i 重新编码，保留 $f(x)$ 中评分最高和次高的维度 y_1 和 y_2（第 2～4 行）；基于阈值 H_{us} 进行筛选，将评估指标 $\varsigma(x_j)$ 小于该阈值的输入定义为不稳定点 $\bar{x}_j$ 加入不稳定集合 $\bar{X}$（第 5～6 行）；之后按照 $\varsigma(x_j)$ 由小到大的顺序对 $\bar{X}$ 中的元素 $\bar{x}_j$ 进行排序，返回新的待验证输入集合（第 7～9 行）.

阶段 2 展示了验证过程，待验证模型 f 以及性质 P 由验证工具编码，得到验证查询 V；其中性质约束 P 由输入和输出驱动，所选择的输入点 $\bar{x}$ 由输入约束 P_{in} 扩展为给定鲁棒半径 δ 下的一个邻域；使用排序后的不稳定集 $\bar{X}$ 替代原始更大规模的输入集 X 提交给验证工具（第 11 行），并通过输出约束 P_{out} 判断性质是否满足，以此来完成对 DNN 的鲁棒性验证（第 12～14 行）. 若大部分 $\bar{x} \in X$ 是安全的，则可以判定该模型的鲁棒性较强（第 15～16 行）；此时我们实现了对不满足性质概率较高的点的优先验证，通过舍弃原始输入集 X 中其他输入点提高验证效率，避免了全部验证的计算开销；若大部分输入点是不安全的，则需要进一步调整阈值 H_{us}（第 17～19 行），从而选取更多的输入点对模型的鲁棒性进行进一步分析. 另外，对于不同类型的验证工具在判断性质是否成立方面会有所差异，我们在下一节讨论了一些集成细节.

4.4 验证工具集成

由于局部鲁棒性的定义存在一定的局限性，导致了对单一输入点的局部验证不具备说服力，因此待验证的输入集 X 中通常需要包含大量的输入点. 我们专注于选择邻域内更有可能出现误分类情况的输入，从而实现局部代表全局的验证. 本文优化方案可以作为一个辅助模块与任何类型的验证工具集成，且具有不同的集成策略和意义.

首先对于完备验证而言，这些工具通常在验证精度方面表现良好，但在验证具有一定规模的神经网络对于一个输入点的局部鲁棒性时可能会花费数千秒甚至更多的时间，并且验证效率会进一步随着网络规模的增加而显著降低，不能胜任较大批量输入数据的验证. 在测试 DNN 的鲁棒性时，一旦发现违规可以立即停止，这使得对中心点进行优先级排序，从最有可能发生冲突的输入域开始，可以有效减少执行时间. 我们通过验证前的输入预分析，基于输出值大小来寻找是否存在距离决策边界相对较近的输入点，并首先关注此类数据点对应的鲁棒邻域，从而实现这种优先级划分. 因此优先次序模式对于这类验证工具有重要意义. 另外，完备的验证方法不仅能够得到确定的验证结果，还能在输入点不满足性质时输出一个反例，而本文方法能够选择这些点并促进反例的生成.

不完备验证通过对模型的近似处理来实现更高的效率，是可扩展验证的主要发展方向. 而过度的近似会导致误差积累从而影响验证精度. 虽然这些方法有较高的效率，但是随着神经网络结构的复杂化同样面临昂贵的计算开销，因此本文的优化思路对于此类工具也同样适用，并且能够预先评估相对安全或不安全的输入以及设置更合理的约束 P_{in}. 另外，这类方法[17]在验证失败时无法进一步对性质成立与否进行分析，这在验证框架中体现在只能给出不可验证（Unknown）的结果而非确定的 Unsafe 结果. 虽然它们无法直接生成具体的反例，但我们选取的不稳定点可以结合对抗样本生成[40-42]以达到更高的成功率，同样有助于有效地训练鲁棒模型（如基于 PGD[32]和 DiffAI[43]的防御训练），这也是本文评估

的重要作用之一.

当验证 DNN 对某一输入点的局部鲁棒性质时,精度较低的工具可能在一个较小的鲁棒半径 δ 下给出不安全或未被成功验证的结果,而高精度工具能验证更大的 δ 值. 此时我们期望 USPs 在各种工具中能验证的最大鲁棒半径 δ_{max} 低于正常样本,因为一些不完备的验证即使性质成立也可能无法给出确定的保证,而该指标能在精度受限时反映神经网络在某个输入空间 η 上的鲁棒性. 对于鲁棒性较差的模型,验证工具的最大可验证 δ 普遍较小,这在本文的鲁棒性评估中对应各输出中存在很多 y_i 与 y_1 值相对接近的情况. 若在给定 δ 下 USPs 满足安全性质,则大致可以判断该模型是相对可靠的. 因此在集成本文优化方法时还需考虑验证工具自身精度及效率、所支持的模型和数据类型以及模型的鲁棒性强弱等因素. 我们在实验中以两种具有代表性的验证工具为例,对集成输入预分析模块后的效果进行了测试.

5　实验评估

在本节中,我们评估了基于输出单元值的鲁棒性评估和基于优先次序的验证,包括决策边界分析的测试证明、不稳定点选择准确率以及集成后对验证工具的效率提升.

5.1　实验设置

严格的形式化技术能够证明本文鲁棒性评估的有效性. 一方面只有通过形式化验证证明模型对各个输入点的鲁棒性才具有更强的理论性和说服力. 另一方面,将基于优先次序的优化模块与验证工具集成也是主要目的之一. 从决策边界分析、鲁棒性评估有效性、集成输入预分析后的运行时间三个角度对所提方法进行了测试.

(1)验证工具

根据不同的验证类型集成了两种先进的验证工具,即基于 SMT 的 Marabou[12,25] 和基于边界传播的 $\beta-$Crown[24],其中 Marabou 是完备的验证工具,而 $\beta-$Crown 支持完备和不完备两种模式. 两种工具均在 2021 年 DNN 验证大赛[44] 中表现优秀,虽然在性能方面取得了较高的得分,但仍无法避免由 DNN 模型复杂度和随机选取大量测试点引起的高验证开销. 我们基于不同输入样本在验证中的判定结果证明 USP 选择的准确率,以此来评价优先次序模式对于验证效率提升的合理性. 在验证中我们默认没有被成功验证的点(超时情况)是相对不安全的.

(2)模型和数据集

考虑到目前验证工具所支持的数据类型、模型结构及激活函数,本节分别使用验证中常用的 ACAS Xu[45]、MNIST 和 CIFAR10 作为实验数据集. 我们在 ACAS Xu 模型[12] 中选取了一个 6×50 结构的神经网络,并在 MNIST 数据集上训练了一个 9×200 的 FNN. 而在 CIFAR10 模型方面我们使用工具中包含的一个验证中常用的基于 PGD 防御训练的 CNN[18]. 以上模型均使用大多数验证工具关注的 ReLU 激活函数.

(3)安全性质

性质规约方面我们使用基于 L_∞ 范数度量的局部鲁棒性,由鲁棒半径 δ 实现参数化,表示扰动输入 x' 中的各特征分量 x'_d(在图像中对应每个像素)与原始输入 x 中的对应特征分量 x_d 的最大距离. 由于不同模型的可靠性以及验证工具的精度存在差异,导致可验证的最大鲁棒半径也有所不同,因此对于每个基准的 δ 有不同设定. 值得注意的是,鲁棒性验证主要关注在神经网络中被正确分类的原始样本. 实验中所采用的验证基准如表 2 所示.

表 2　验证基准

工具	完备	模型	结构	数据集	鲁棒半径
Marabou	√	FNN	6×50	ACAS Xu	0.025 0.075
$\beta-$Crown	×	FNN	9×200	MNIST	0.015 0.018
$\beta-$Crown	√	CNN	Conv	CIFAR10	2/255

5.2　实验结果

(1)突变测试与决策边界分析. 首先通过突变测试来证明 DB 分析的正确性. 以 MNIST 中的 5 个输入点为例展示了基于模型输出值大小的度量指标 ς 与突变模型中错分类频率 ς 之间的关系,结果如表 3 所示. 在 MT 中模型的生成具有随机性,为了提高评估准确性,这里我们分别使用 GF 和 NAI 算子以更高的突变率 R_M 生成足够数量的 MMs.

表 3　基于输出值与 MT 的鲁棒性评估结果比较

x_j	c_1	c_2		c_3		M_2	M_3	判定
x_1	2	7	7	1	1	99/500	50/500	不稳定点
x_2	7	9	9	3	3	41/500	30/500	正常样本
x_3	8	3	3	2	2	82/500	43/500	不稳定点
x_4	2	7	7	8	1	31/500	18/500	正常样本
x_5	3	9	9	5	2	68/500	62/500	不稳定点

注:其中 c_1 为正确标签,c_2 和 c_3 分别为第二和第三大输出值 y_2 和 y_3 所对应的标签(左),以及 MMs 中产生次数最多的两个错分类标签(右);M_2 和 M_3 分别对应 c_2 和 c_3 的出现频率 N_{c_i}/N_M,且 MMs 数量 N_M 为 500.

突变测试中的出错次数能在很大程度上度量多分类模型中输入点与不同错分类标签对应的决策边界之间的距离关系.考虑到局部规约特点,我们优先关注与输入点距离更近的边界,所以只取了第二大和第三大输出值 y_2 和 y_3.从表 2 可得,在原始模型中输出值较大的节点对应的标签 c_2 和 c_3 与突变模型中出现频率最高的两个错分类标签基本一致.指标 $\varsigma(x)$ 的大小由满足阈值 H_{us} 条件的输出值 y_i 决定,y_i 与突变测试中的 M 值成正相关,且 x 往往都距离 y_2 对应的错分类 DB 更近.因此 $\varsigma(x)$ 同样能够反映输入点 x 与 DB 间的距离关系.另外,若 $f(x)$ 中存在多个与最大输出分量相差较小的维度,可以认为该输入点 x 同时位于多条 DBs 附近,这在 MMs 的输出结果中同样能够体现(例如 x_5),此时在鲁棒性分析中优先关注距离最近的那条即可,即在 MMs 中误判次数最多的标签 $c_2=9$.

(2)不稳定点选择准确率.通过验证工具对所选样本进行测试,以证明神经网络对所定位的 DB 附近的点更有可能是不鲁棒的.测试中的两种工具都具有一定的精度基础,能够为 DNN 的局部鲁棒性提供较为精确的验证.考虑到验证工具自身的性能使结果便于统计,对各基准分别选取了 30、100、50 个测试样本,测试了不稳定点在验证中违反性质的概率,并在不完备验证中增加了一项平均可验证的最大鲁棒半径 δ_{max} 的对比.为了统一评价标准,基于测试样本的输出差值对阈值 H_{us} 进行归一化处理,通过 H_{us} 将 N_{us} 控制在总输入数量的 50%和 30%.在给定鲁棒半径下,若 USPs 更容易违反安全性质则可以表明鲁棒性度量指标的有效性,结果如表 4、表 5 所示.

表 4 完备验证中不稳定点测试结果

	H_{us}	δ	N_{us}	A_1	A_2
$M-1$	0.51	0.025	15	33.3%	100.0%
		0.075		80.0%	92.3%
$M-2$	0.34	0.025	9	55.6%	100.0%
		0.075		100.0%	69.2%
$\beta C-3$	0.47	2/255	25	92.0%	85.2%

注:验证结果包含了安全和不安全的点;A_1 表示不稳定集内不安全数量的比例,A_2 表示不稳定集内不安全数量在所有违反性质数量中的比例.

由表 4 和表 5 可得,在给定鲁棒半径下不稳定点有更高的不鲁棒概率,并且由 N_{us} 的设置变化可以看出,ς 值越小,验证结果为不安全的概率就越大.而模型在某一输入点上的鲁棒性越差说明了对抗攻击所需的扰动越小以及该输入距离 DB 较近.对于完备的验证,每得到一个输入不满足性质时会发现一个反例.在不完备的验证中,不可验证的点是相对不安全的,虽然无法对这些点的鲁棒性做进一步分析,但是由 $\varsigma(x)$ 值判定为较稳定的输入点大概率是可验证的,以及 USPs 的平均 δ_{max} 低于正常样本两个方面也能够证明鲁棒性评估是有效的.而近似的不完备方法的优势是可以处理规模较大的复杂网络,此时可以侧重于大规模输入数据预筛选带来的效率提升.

表 5 不完备验证中不稳定点测试结果

	H_{us}	δ	N_{us}	A_1	A_2	δ_{max}
$\beta C-1$	0.57	0.015	50	60.0%	96.8%	0.022 0.012
		0.018		86.0%	86.0%	
$\beta C-2$	0.41	0.015	30	83.3%	80.6%	0.020 0.009
		0.018		96.7%	58.0%	

注:验证结果包含了已验证和不可验证的点;A_1 和 A_2 与表 4 设定相同;最后一列分别展示了普通样本(左)和不稳定点(右)平均可验证的鲁棒半径.

由 USPs 在验证工具中较高的不安全概率表明本文方法也可以单独作为一种 DNN 模型的鲁棒性分析方法使用.虽然无法直接给出神经网络可接受的扰动,但如果一个模型中存在较多最大输出分量与其他输出分量差值相对较小的点,那么该模型大概率存在较为严重的鲁棒性隐患,而不仅仅是像不完备的验证一样只能给出不可验证的结果.形式化方法基于严格的数学推导,其精确的逻辑论证是不可替代的,因此在集成输入预分析模块时,即使结果中可能存在一些精度误差,但是不会影响模型整体的鲁棒性分析,尤其对于包含大批量输入的验证任务.

(3)验证效率提升.分别以完备的 Marabou 和不完备 β-Crown 验证基准为例说明优先次序优化对所集成工具的效率提升.基于不稳定点更容易不满足鲁棒性这一实验结果,通过只验证 USPs 实现对一个 DNN 模型的鲁棒性证明.待验证的高优先级任务数量由可调整的不稳定阈值决定.

我们统计了集成优先次序框架后的验证时间变化,如图 4 所示.由于本文优化主要通过输入点的筛选以减少待验证查询的数量,所以验证效率显然取得了显著的提升.这种验证模式的开销主要包括常规的验证查询处理以及输入的鲁棒性预分析,其中验证中约束求解的效率取决于验证工具本身的编码和计算方式;在 Marabou 中可以观察到该工具在验证不安全样本时的效率更高;但 β-Crown 不完备模式的效率会受到超时的限制;而完备模式中的验

证时间减少将更加明显.在预分析方面只需执行多次前向传播,不会对验证算法本身的性能产生影响,这是据我们所知最轻量级的评估.另外对于待验证的 DNN 模型而言,在 USPs 的选择中虽然不能达到完全准确,但能够过滤掉大量低验证需求的输入.即使是对于效率更低的高精度和大规模基准,也可以通过只验证相对少量的 USPs 对模型整体的可靠性进行评估.

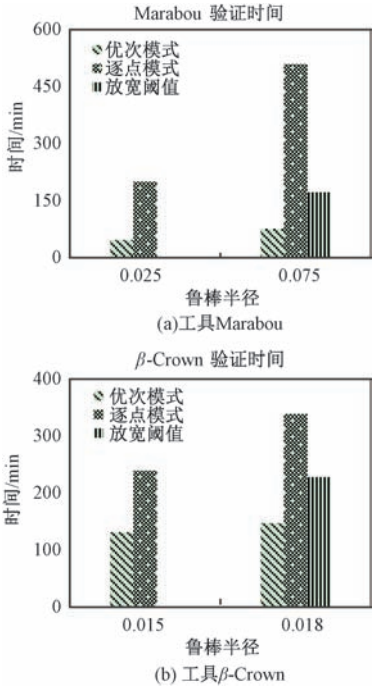

图 4　优化前后的运行时间对比

综上所述,在目前的 DNN 鲁棒性验证中人们更希望了解一个神经网络对大量输入的鲁棒性,而不仅仅是对某单一样本是否可靠的简单答案.我们的优先次序方法能够对理论上存在的任意合理输入进行预分析,从而避免了随机选取测试点验证的说服力不足,节省了验证神经网络在大批量输入点上的局部鲁棒性的昂贵计算开销,同时还能相对准确地分析模型的鲁棒性缺陷.需要验证的输入数量越多,这种优化就越有意义,因此在结合使用时不仅减少了传统逐点验证中的计算消耗,并且在很大程度上提高了局部验证的说服力.

6　总结与展望

本文提出了一种基于神经网络输出层单元值的鲁棒性评估方法.该方法根据距离决策边界越近的点,其鲁棒性越弱,且对应的输出分量往往越大,从而评估模型在不同输入样本上的局部鲁棒性.针对目前神经网络鲁棒性验证中的效率缺陷,提出了一种优先次序的优化方案与现有的验证工具进行集成.通过对传统的验证模式进行改进,基于鲁棒性度量指标实现对不同输入点验证优先级的划分,以此来舍弃低验证需求样本并充分分析模型中的鲁棒性问题.对所提方法进行了实现和测试,通过对比基于鲁棒性评估选择的不稳定点在形式化验证中结果的一致性以及集成输入预分析模块后对验证效率带来的影响,证明了该方法的有效性.

本文方法能高效地评估模型对不同输入的鲁棒性,且方法本身不受验证中的模型类型、应用领域以及可扩展性问题的限制.虽然在结果的准确性和说服力方面不如形式化方法,但是它在一定程度上提供了一个当安全性质可能被侵犯时关于网络有多可靠的信息概念[46].而且这种鲁棒性评估在与更严格的验证工具集成使用时能够取得良好的互补,尤其对于难以兼顾效率的高精度的验证工具.因此,通过选择具有代表性的点优先验证来实现对网络的可靠性分析是一个可行的优化路线.随着验证技术的发展能够支持更多的应用场景、数据类型以及复杂结构和规模的神经网络时,本文方法会有更高的实际应用价值.此外,本文基于鲁棒性问题分析,从决策边界角度探索了深度学习模型的可解释性.

对于 DNN 可靠性保证的研究,未来工作将主要围绕以下几个方向展开:

(1)研究 DNN 输出值与性质之间的量化关系,如决策边界距离和扰动大小的定量分析,并探索其他安全性质层面的刻画,从而为模型的可靠性提供更精确的度量.

(2)研究 DNN 模型的高层次抽象编码方法,在建模层面对复杂神经网络的编码进行简化,提高以网络自身作为验证模型的验证效率和可扩展性.

(3)研究其他面向 DNN 鲁棒性验证的性能优化方法,例如基于并行优化通过投入更多的硬件资源,以缓解现有的形式化验证算法难以完全克服的效率问题.

(4)研究 DNN 鲁棒性验证技术在其他模型结构和非图像分类领域中的应用,以探索并观察验证技术在不同类型的模型及数据中的表现.

参　考　文　献

[1] Samek W, Montavon G, Lapuschkin S, et al. Explaining deep neural networks and beyond: A review of methods and applica-

tions. Proceedings of the IEEE, 2021, 109(3): 247-278

[2] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks. Communications of the ACM, 2017, 60(6): 84-90

[3] Graves A, Mohamed A, Hinton G. Speech recognition with deep recurrent neural networks//Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing. Vancouver, Canada, 2013: 6645-6649

[4] Goldberg Y. Neural network methods in natural language processing (synthesis lectures on human language technologies). San Rafael, USA: Morgan & Claypool, 2017

[5] Bojarski M, Del Testa D, Dworakowski D, et al. End to end learning for self-driving cars. arXiv preprint arXiv: 1604.07316, 2016

[6] Xu Meng-Ting, Zhang Tao, Li Zhong-Nian, et al. Towards evaluating the robustness of deep diagnostic models by adversarial attack. Medical Image Analysis, 2021, 69(1): 101977

[7] Arp D, Spreitzenbarth M, Hubner M, et al. Drebin: Effective and explainable detection of android malware in your pocket//Proceedings of the Network and Distributed System Security Symposium. San Diego, USA, 2014, 14: 23-26

[8] Szegedy C, Zaremba W, Sutskever I, et al. Intriguing properties of neural networks//Proceedings of the International Conference on Learning Representations. Banff, Canada, 2014: 1-10

[9] Goodfellow I J, Shlens J, Szegedy C. Explaining and harnessing adversarial examples//Proceedings of the International Conference on Learning Representations. San Diego, USA, 2015: 1-11

[10] Pulina L, Tacchella A. An abstraction-refinement approach to verification of artificial neural networks//Proceedings of the International Conference on Computer Aided Verification. Edinburgh, UK, 2010: 243-257

[11] Henriksen P, Lomuscio A. Efficient neural network verification via adaptive refinement and adversarial search//Proceedings of the European Conferenceon Artificial Intelligence. Santiago de Compostela, Spain, 2020: 2513-2520

[12] Katz G, Barrett C, Dill D L, et al. Reluplex: An efficient SMT solver for verifying deep neural networks//Proceedings of the International Conference on Computer Aided Verification. Heidelberg, Germany, 2017: 97-117

[13] Ehlers R. Formal verification of piece-wise linear feed-forward neural networks//Proceedings of the International Symposium on Automated Technology for Verification and Analysis. Pune, India, 2017: 269-286

[14] Tjeng V, Xiao Kai, Tedrake R. Evaluating robustness of neural networks with mixed integer programming//Proceedings of the International Conference on Learning Representations. New Orleans, USA, 2019: 1-21

[15] Xiang Wei-Ming, Tran H-D, Johnson T T. Output reachable set estimation and verification for multilayer neural networks. IEEE Transactions on Neural Networks and Learning Systems, 2018, 29(11): 5777-5783

[16] Wang Shi-Qi, Pei Ke-Xin, Whitehouse J, et al. Efficient formal safety analysis of neural networks//Proceedings of the International Conference on Neural Information Processing Systems. Montreal, Canada, 2018: 6367-6377

[17] Singh G, Gehr T, Püschel M, et al. An abstract domain for certifying neural networks. Proceedings of the ACM on Programming Languages. 2019, 3(POPL): 1-30

[18] Singh G, Ganvir R, Püschel M, et al. Beyond the single neuron convex barrier for neural network certification//Proceedings of the International Conference on Neural Information Processing Systems. Vancouver, Canada, 2019: 15072-15083

[19] Wong E, Kolter Z. Provable defenses against adversarial examples via the convex outer adversarial polytope//Proceedings of the International Conference on Machine Learning. Stockholmsmssan, Sweden, 2018: 5286-5295

[20] WengTsui-Wei, Zhang Huan, Chen Hong-Ge, et al. Towards fast computation of certified robustness for ReLU networks//Proceedings of the International Conference on Machine Learning. Stockholmsmssan, Sweden, 2018: 5276-5285

[21] Zhang Huan, WengTsui-Wei, Chen Pin-Yu, et al. Efficient neural network robustness certification with general activation functions//Proceedings of the International Conference on Neural Information Processing Systems. Montreal, Canada, 2018: 4944-4953

[22] WengTsui-Wei, Zhang Huan, Chen Pin-Yu, et al. On extensions of clever: A neural network robustness evaluation algorithm//Proceedings of the IEEE Global Conferenceon Signal and Information Processing. Anaheim, USA, 2018: 1159-1163

[23] Katz G, Barrett C, Dill D L, et al. Towards proving the adversarial robustness of deep neural networks//Proceedings of the Workshop on Formal Verification of Autonomous Vehicles, Turin, Italy, 2017: 19-26

[24] Wang Shi-Qi, Zhang Huan, Xu Kai-Di, et al. Beta-crown: Efficient bound propagation with per-neuron split constraints for neural network robustness verification//Proceedings of the International Conference on Neural Information Processing Systems. Montreal, Canada, 2021: 29909-29921

[25] Katz G, Huang D A, Ibeling D, et al. The marabou framework for verification and analysis of deep neural networks//Proceedings of the International Conference on Computer Aided Verification. New York, USA, 2019: 443-452

[26] Feng Cheng-Dong, Chen Zhen-Bang, Hong Wei-Jiang, et al. Boosting the robustness verification of DNN by identifying the Achilles's heel. arXiv preprint arXiv: 1811.07108, 2018

[27] WengTsui-Wei, Zhang Huan, Chen Pinyu, et al. Evaluating the robustness of neural networks: An extreme value theory approach//Proceedings of the International Conference on Learning Representations. Vancouver, Canada, 2018: 1-18

[28] Yousefzadeh R, O'Leary D P. Investigating decision boundaries of trained neural networks. arXiv preprint arXiv:

1908.02802, 2019

[29] Elsayed G, Krishnan D, Mobahi H, et al. Large margin deep networks for classification//Proceedings of the International-Conference on Neural Information Processing Systems. Montreal, Canada, 2018: 850-860

[30] Moosavi-Dezfooli S M, Fawzi A, Frossard P. Deepfool: a simple and accurate method to fool deep neural networks//Proceedings of the IEEE Conferenceon Computer Vision and Pattern Recognition. Las Vegas, USA, 2016: 2574-2582

[31] Selvaraju R R, Cogswell M, Das A, et al. Grad-cam: Visual explanations from deep networks via gradient-based localization//Proceedings of the IEEE International Conference on Computer Vision. Venice,Italy, 2017: 618-626

[32] Madry A, Makelov A, Schmidt L, et al. Towards deep learning models resistant to adversarial attacks//Proceedings of the International Conference on Learning Representations. Vancouver, Canada, 2018: 1-28

[33] Jin Hai-Bo, Chen Jin-Yin, Zheng Hai-Bin, et al. ROBY: Evaluating the robustness of a deep modelby its decision boundaries. Information Sciences, 2022, 587(1): 97-122

[34] Mickisch D, Assion F, Greßner F, et al. Understanding the decision boundary of deep neural networks: An empirical study. arXiv preprint arXiv:2002.01810, 2020

[35] Wang Jing-Yi, Dong Guo-Liang, Sun Jun, et al. Adversarial sample detection for deep neural network through model mutation testing//Proceedings of the International Conference on Software Engineering. Montreal, Canada, 2019: 1245-1256

[36] Ma Lei, Zhang Fu-Yuan, Sun Ji-Yuan, et al. Deepmutation: Mutation testing of deep learning systems//Proceedings of the IEEE International Symposium on Software Reliability Engineering. Memphis, USA,2018: 100-111

[37] Lin Ren-Hao, Zhou Qing-Lei, Wu Bin, et al. Robustness evaluation for deep neural networks via mutation decision boundaries analysis. Information Sciences, 2022, 601(1): 147-161

[38] LecunY ,Bottou L . Gradient-based learning applied to document recognition. Proceedings of the IEEE, 1998, 86(11): 2278-2324

[39] Wang Lin, Yang Zheng-Feng, Chen Xin, et al. Robustness verification of classification deep neural networks via linear programming//Proceedings of the IEEE Conferenceon Computer Vision and Pattern Recognition. Long Beach, USA, 2019: 11418-11427

[40] Kurakin A, Goodfellow I J, Bengio S. Adversarial examples in the physical world//Proceedings of the International Conference on Learning Representations. Toulon, France, 2017: 1-14

[41] Carlini N, Wagner D. Towards evaluating the robustness of neural networks//Proceedings of the IEEE Symposium on Security and Privacy. San Jose, USA, 2017: 39-57

[42] Dong Yin-Peng, Liao Fang-Zhou, Pang Tian-Yu, et al. Boosting adversarial attacks with momentum//Proceedings of the IEEE Conferenceon Computer Vision and Pattern Recognition. Salt Lake City, USA, 2018: 9185-9193

[43] Mirman M, Gehr T, Vechev M. Differentiable abstract interpretation for provably robust neural networks//Proceedings of the International Conference on Machine Learning. Stockholmsmssan, Sweden, 2018: 3578-3586

[44] Bak S, Liu Chang-Liu, Johnson T. The second international verification of neural networks competition (vnn-comp 2021): Summary and results. arXiv preprint arXiv: 2109.00498, 2021

[45] Julian K D, Lopez J, Brush J S, et al. Policy compression for aircraft collision avoidance systems//Proceedings of the Digital Avionics Systems Conference. Sacramento, USA, 2016: 1-10

[46] Webb S, Rainforth T, Teh Y W, et al. A statistical approach to assessing neural network robustness//Proceedings of the International Conference on Learning Representations. New Orleans, USA, 2019: 1-15

LIN Ren-Hao, Ph. D. candidate. His main research interests include AI security and information security.

ZHOU Qing-Lei, Ph. D., professor. His main research interests include formal method and AI algorithm.

HU Tian-Qing, Ph. D. candidate. Her main research interests include image processing and machine learning.

WANG Yi-Feng, Ph. D. candidate. His main research interests include network security and machine learning.

Background

This paper is the frontier research in the field of AI security and DNN verification. Deep neural network is the most effective deep learning model. However, DNN models usually suffer from serious robustness threats, where imperceptible perturbations in the input samples can change the decision of DNN models. The inherent defect of robustness is likely to bring incalculable losses to people in safety-critical systems, so the trustworthiness assurance technology of DNN model has become an important and cutting-edge research at home and abroad in recent years. Among them, formal verification technology is based on strict mathematical derivation, which is a reliable means to prove the security properties of DNN

models. Existing verification techniques mainly use DNN itself as the verification model, but its nonlinear and large-scale characteristics lead to serious efficiency and scalability problems in DNN verification techniques.

Whether for complete or incomplete verification tools, how to balance verification efficiency and precisionis a challenge. Even the state-of-the-art verification tools in the 2021 DNN verification competition face this problem. Current verification algorithms are becoming saturated, and still cannot completely overcome the efficiency caused by high computational overhead. Therefore, it is necessary to design appropriate performance optimization methods for DNN verification. Some of them include parallelization of GPU mode and branch and bound, which mainly improve verification efficiency by investing more computational resources.

This paper is dedicated to making verification technology better applied in practice, focusing on a novel optimization idea: prioritization. Considering the limitations of the local robustness specification, the traditional point-by-point verification mode is replaced by selecting relatively unsafe inputs. It can greatly improve the efficiency of verification by reducing the number of tasks to be verified while ensuring a relatively accurate analysis of the robustness of the model. This is an optimization scheme that has hardly been used in current verification techniques. Moreover, based on the decision boundary analysis theory, the robustness of DNN models is discussed in detail. This research group has been engaged in research of formal methods and deep learning, and has proposed a robustness evaluation for DNN models based on mutation testingin DNN prioritization verification. This paper investigates a more lightweight robustness evaluation to guide the selection of points for prioritization verification. With the development of verification technology, the optimization idea in this paper may make greater contributions to the practical application of DNN verification.