深度学习模型鲁棒性研究综述

纪守领[1] 杜天宇[1] 邓水光[1] 程鹏[2] 时杰[3] 杨珉[4] 李博[5]

[1](浙江大学计算机科学与技术学院, 杭州 310027)

[2](浙江大学控制科学与工程学院, 杭州 310027)

[3](华为新加坡研究所, 新加坡 138589)

[4](复旦大学计算机科学与技术学院, 上海 201203)

[5](伊利诺伊大学香槟分校计算机科学学院, 厄巴纳香槟 美国 61822)

摘 要 在大数据时代下，深度学习理论和技术取得的突破性进展，为人工智能提供了数据和算法层面的强有力支撑，同时促进了深度学习的规模化和产业化发展。然而，尽管深度学习模型在现实应用中有着出色的表现，但其本身仍然面临着诸多的安全威胁。为了构建安全可靠的深度学习系统，消除深度学习模型在实际部署应用中的潜在安全风险，深度学习模型鲁棒性分析问题吸引了学术界和工业界的广泛关注，一大批学者分别从精确和近似的角度对深度学习模型鲁棒性问题进行了深入的研究，并且提出了一系列的模型鲁棒性量化分析方法。在本综述中，我们回顾了深度学习模型鲁棒性分析问题当前所面临的挑战，并对现有的研究工作进行了系统的总结和科学的归纳，同时明确了当前研究的优势和不足，最后探讨了深度学习模型鲁棒性研究以及未来潜在的研究方向。

关键词 深度学习；对抗样本；鲁棒性分析；人工智能安全

中图法分类号

Robustness Certification Research on Deep Learning Models: A Survey

JI Shou-Ling[1] DU Tian-Yu[1] DENG Shui-Guang[1] CHENG Peng[2] SHI Jie[3] YANG Min[4] LI Bo[5]

[1](College of Computer Science and Technology, Zhejiang University, Hangzhou 310027, China)

[2](College of Control Science and Engineering, Zhejiang University, Hangzhou 310027)

[3](Huawei Singapore Research Center, Singapore 138589)

[4](School of Computer Science, Fudan University, Shanghai 201203)

[5](Department of Computer Science, University of Illinois at Urbana–Champaign, Urbana-Champaign 61822, United States)

Abstract In the era of big data, breakthroughs in theories and technologies of deep learning have provided strong support for artificial intelligence at the data and the algorithm level, as well as have promoted the development of scale and industrialization of deep learning in a large number of tasks, such as image classification, object detection, semantic segmentation, natural language processing and speech recognition. However, though deep learning models have excellent performance in many real-world applications, they still suffer many security threats. For instance, it is now known that deep neural networks are fundamentally vulnerable to malicious manipulations, such as

本课题得到浙江省自然科学基金杰出青年项目(LR19F020003)、国家重点研发计划项目(2020YFB2103802)、国家自然科学基金项目(61772466, U1936215, U1836202)、中央高校基本科研业务费专项资金（浙江大学NGICS大平台）资助.纪守领(通信作者)，男，1986年生，博士，研究员，计算机学会(CCF)会员（66979M）,主要研究领域为人工智能与安全、数据驱动安全、软件与系统安全、大数据分析.E-mail: sji@zju.edu.cn. 杜天宇，女，1996年生，博士研究生，计算机学会(CCF)学生会员（9603239）,主要研究领域为人工智能安全.E-mail: zjradty@zju.edu.cn. 邓水光，男，博士，教授，计算机学会(CCF)会员（06679S）,主要研究领域为服务计算、边缘计算、流程管理和大数据等.E-mail: dengsg@zju.edu.cn. 程鹏，男，博士，教授，计算机学会(CCF)会员（10568M），主要研究领域为控制系统安全、物联网/信息物理融合系统、数据安全与隐私保护.E-mail: pcheng@iipc.zju.edu.cn. 时杰，男，博士，主要研究领域为人工智能安全与隐私.E-mail: shi.jie1@huawei.com. 李博，女，博士，助理教授，主要研究领域为人工智能安全、博弈论等.E-mail: lbo@illinois.edu.

adversarial examples that force target deep neural networks to misbehave. In recent years, a plethora of work has focused on constructing adversarial examples in various domains. The phenomenon of adversarial examples demonstrates the inherent lack of robustness of deep neural networks, which limits their use in security-critical applications. In order to build a safe and reliable deep learning system and eliminate the potential security risks of deep learning models in real-world applications, the security issue of deep learning has attracted extensive attention from academia and industry. Thus far, intensive research has been devoted to improving the robustness of DNNs against adversarial attacks. Unfortunately, most defenses are based on heuristics and thus lack any theoretical guarantee, which can often be defeated or circumvented by more powerful attacks. Therefore, defenses only showing empirical success against attacks, are difficult to be concluded robust. Aiming to end the constant arms race between adversarial attacks and defenses, the concept of robustness certification is proposed to provide guaranteed robustness by formally verifying whether a given region surrounding a data point admits any adversarial example. Robustness certification, the functionality of verifying whether the given region surrounding a data point admits any adversarial example, provides guaranteed security for deep neural networks deployed in adversarial environments. Within the certified robustness bound, any possible perturbation would not impact the prediction of a deep neural network. A large number of researchers have conducted in-depth research on the model robustness certification from the perspective of complete and incomplete, and proposed a series of certification methods. These methods can be generally categorized as exact certification methods and relaxed certification methods. Exact certification methods are mostly based on satisfiability modulo theories or mixed-integer linear program solvers. Though these methods are able to certify the exact robustness bound, they are usually computationally expensive. Hence, it is difficult to scale them even to medium size networks. Relaxed certification methods include the convex polytope methods, reachability analysis methods, and abstract interpretation methods, etc. These methods are usually efficient but cannot provide precise robustness bounds as exact certification methods do. Nevertheless, considering the expensive computational cost, relaxed certification methods are shown to be more promising in practical applications, especially for large networks. In this survey, we review current challenges of model robustness certification problem, systematically and scientifically summarize existing research work, and clarify the advantages and disadvantages of current research. Finally, we explore future research directions of model robustness certification research.

Key words deep learning; adversarial example; robustness certification; artificial intelligence security

1 引言

受益于计算力和智能设备的飞速发展，全世界正在经历第三次人工智能浪潮。人工智能以计算机视觉、序列处理、智能决策等技术为核心在各个应用领域展开，并延伸到人类生活的方方面面，包括自适应控制[1]、模式识别[2,118]、游戏[3]以及自动驾驶[4]等安全攸关型应用。例如，无人驾驶飞机防撞系统（Aircraft Collision Avoidance System, ACAS）使用深度神经网络根据附近入侵者飞机的位置和速度来预测最佳行动。然而，尽管深度神经网络已经显示出解决复杂问题的有效性和强大能力，但它们仅限于仅满足最低安全完整性级别的系统，因此它们在安全关键型环境中的采用仍受到限制，主要原因在于在大多数情况下神经网络模型被视为无法对其预测行为进行合理解释的黑匣子，并且在理论上难以证明其性质[117]。

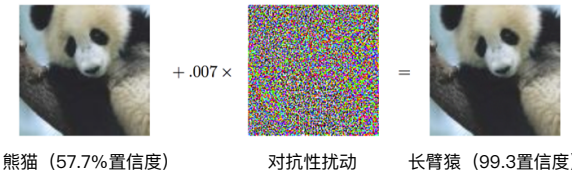


图 1 对抗样本示例[5-11]。

随着深度学习的对抗攻击领域日益广泛，对抗样本的危险性日益凸显[7,12,13,109,110,112,114-116]，即通过向正常样例中添加精细设计的、人类无法感知的扰动达到不干扰人类认知却能使机器学习模型做出错误判断。以图像分类任务为例，如图 1 所示，原始样本以 57.7%的置信度被模型分类为“熊猫”,而

添加对抗扰动之后得到的样本则以 99.3%的置信度被错误地分类为“长臂猿”，然而对于人而言，对抗样本依然会被视为熊猫。由于这种细微的扰动通常是人眼难以分辨的，因而使得攻击隐蔽性极强、危害性极大，给 ACAS 等安全攸关型应用中部署的深度学习模型带来了巨大的安全威胁。

为了防御对抗样本攻击，研究者进行了一系列的防御方法探索[5-11,113]。然而，即使是被广泛认可并且迄今为止最成功的ℓ_∞防御[5]，它的ℓ_0鲁棒性比未防御的网络还低，并且仍然极易受到ℓ_2的扰动影响[14,111]。这些结果表明，仅对对抗攻击进行经验性的防御无法保证模型的鲁棒性得到实质性的提升，模型的鲁棒性需要一个定量的、有理论保证的指标进行评估。因此，如果要将深度学习模型部署到诸如自动驾驶汽车等安全攸关型应用中，我们需要为模型的鲁棒性提供理论上的安全保证，即计算模型的鲁棒性边界。模型鲁棒性边界是针对某个具体样本而言的，是保证模型预测正确的条件下样本的最大可扰动范围，即模型对这个样本的分类决策不会在这个边界内变化。具体地，令输入样本x的维度为d，输出类别的个数为K，神经网络模型为$f:\mathbb{R}^d \to \mathbb{R}^K$，输入样本的类别为$c = argmax\ f_j(x), j = 1,2,\dots,K$，在$\ell_p$空间假设下，模型对$x$提供$\epsilon$-鲁棒性保证表明模型对$x$的分类决策不会在这个样本$\ell_p$空间周围$\epsilon$大小内变化。

在本文中，我们首先阐述了深度学习模型鲁棒性分析现存的问题与挑战，然后从精确与近似两个角度对现有的鲁棒性分析方法进行系统总结和科学归纳，并讨论了相关研究的局限性。最后，我们讨论了模型鲁棒性分析问题未来的研究方向。

2 问题与挑战

目前，深度神经网络的鲁棒性分析问题的挑战主要集中在以下几个方面：

（1）神经网络的非线性特点。由于非线性激活函数和复杂结构的存在，深度神经网络具有非线性、非凸性的特点，因此很难估计其输出范围，并且验证分段线性神经网络的简单特性也已被证明是 NP 完全问题[15]。这一问题的难点在于深度神经网络中非线性激活函数的存在。具体地，深度神经网络的每一层由一组神经元构成，每个神经元的值是通过计算来自上一层神经元的值的线性组合，然后将激活函数应用于这一线性组合。由于这些激活函数是非线性的，因此这一过程是非凸的。以应用最为广泛的激活函数 ReLU 为例，当 ReLU 函数应用于具有正值的节点时，它将返回不变的值，但是当该值为负时，ReLU 函数将返回 0。然而，使用 ReLU 验证 DNN 属性的方法不得不做出显著简化的假设，例如仅考虑所有 ReLU 都固定为正值或 0 的区域[16]。直到最近，研究人员才能够基于可满足性模理论等形式方法，对最简单的 ReLU 分段线性神经网络进行了初步验证[15,21]。由于可满足性模理论求解器难以处理非线性运算，因此基于可满足性模理论的方法通常只适用于激活函数为分段线性的神经网络，无法扩展到具有其它类型激活函数的神经网络。

（2）神经网络的大规模特点。在实际应用中，性能表现优秀的神经网络通常具有大规模的特点。因此，尽管每个 ReLU 节点的线性区域可以划分为两个线性约束并有效地进行验证，但是由于线性片段的总数与网络中节点的数量成指数增长[17,18]，对整个网络进行精确验证是非常困难的。这是因为对于任何大型网络，其所有组合的详尽枚举极其昂贵，很难准确估计输出范围。此外，基于可满足性模理论的方法严重受到求解器效率的限制，仅能处理非常小的网络（例如，只有 10 到 20 个隐藏节点的单个隐藏层[20]），无法扩展到大多数现实世界中的大型网络，而基于采样的推理技术（例如黑盒蒙特卡洛采样）也需要大量数据才能在决策边界上生成严格的准确边界[19]。

总之，由于不同学者所处的研究领域不同，解决问题的角度不同，所提出的鲁棒性分析方法也各有侧重，因此亟需对现有的研究工作进行系统的整理和科学的归纳、总结、分析。典型的模型鲁棒性分析方法总结如表 1 所示。目前的模型鲁棒性分析方法主要分为两大类：（1）精确方法：可以证明精确的鲁棒性边界，但计算复杂度高，在最坏情况下计算复杂度相对于网络规模是成指数增长的，因此通常只适用于极小规模的神经网络；（2）近似方法：效率高、能够扩展到复杂神经网络，但只能证明近似的鲁棒性边界。

表 1 典型的模型鲁棒性分析方法总结

方法	精确	模型	效率	激活函数	领域
Reluplex[15]	✓	FCN	○	ReLU	图像
PLANET[21]	✓	FCN	○	ReLU	图像
Huang et al.[22]	✓	FCN	○	arbitrary	图像

Tjeng et al.[23]	✓	FCN, CNN	◐	ReLU	图像
Szgedy et al.[12]	×	FCN	●	ReLU	图像
Hein et al.[24]	×	FCN	●	ReLU	图像
SDP[25]	×	FCN	○	ReLU	图像
Wong et al.[26]	×	FCN, CNN	◐	ReLU	图像
Dvijotham[27]	×	FCN	◐	任意	图像
Fast-lin/Fast-lip[28]	×	FCN	●	ReLU	图像
CROWN[29]	×	FCN	●	3 种	图像
CNN-Cert [30]	×	FCN, CNN	●	ReLU	图像
CLEVER[31]	×	任意	●	任意	图像
PixelDP[32]	×	FCN, CNN	◐	任意	图像
Cohen et al.[33]	×	任意	◐	任意	图像
AI^2[34]	×	FCN, CNN	◐	ReLU	图像
DeepZ[35]	×	FCN, CNN	●	3 种	图像
DiffAI[36]	×	FCN, CNN	◐	ReLU	图像
DeepPoly[37]	×	FCN, CNN	●	3 种	图像
RefineAI[38]	×	FCN, CNN	◐	ReLU	图像
k-ReLU[39]	×	FCN, CNN	◐	ReLU	图像
IBP[40]	×	FCN, CNN	◐	ReLU	图像
Chen et al.[41]	×	Tree-based	◐	/	图像
Wang et al.[42]	×	k-nearest	◐	/	图像
Huang et al.[43]	×	CNN	◐	ReLU	文本
POPQORN[44]	×	RNN	◐	2 种	文本
Jia et al.[45]	×	RNN	●	2 种	文本
Shi et al.[46]	×	RNN	◐	2 种	文本
Zügner et al.[47]	×	GNN	◐	ReLU	图
Bojchevski[48]	×	GNN	◐	ReLU	图

注：✓=满足，×=不满足；效率：○=低，◐=中，●=高。

3 精确方法

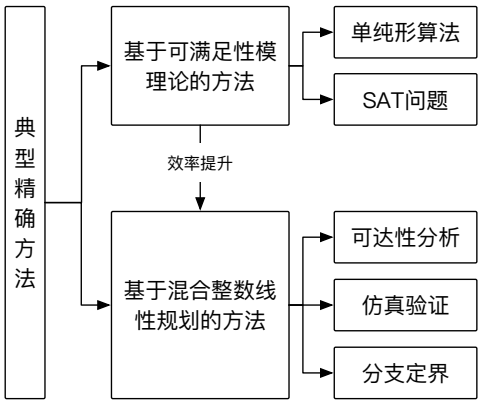


图 2：典型模型鲁棒性精确分析方法概念图。

精确方法主要是基于离散优化（Discrete Optimization）理论来形式化验证神经网络中某些属性对于任何可能的输入的可行性，即利用可满足性模理论（Satisfiability Modulo Theories, SMT）或混合整数线性规划（Mixed Integer Linear Programming, MILP）来解决此类形式验证问题。这类方法通常是通过利用 ReLU 的分段线性特性并在搜索可行解时尝试逐渐满足它们施加的约束来实现的。图 2 梳理了典型模型鲁棒性精确分析方法的相关研究工作。

3.1 基于可满足性模理论的方法

Katz 等人提出具有实数算术原理的 SMT 求解器 Reluplex[15]，通过扩展单纯形（simplex）算法（用于解决 LP 实例的标准算法）以支持 ReLU 约束，并验证了激活函数为 ReLU 的前馈神经网络的鲁棒性。虽然他们使用了线性规划作为与 SMT 的比较，但没有提到对线性规划的任何优化方法。此外，这种方法效率非常低，对于一个约含 100 个神经元的小规模网络来说，计算一个样本至少需要若干个小时。因此，虽然 Reluplex 可以形式上地验证神经网络的一些特性，但由于计算成本太高，无法扩展到实际模型中。此外，Reluplex 通过将 ReLU 网络描述为一个分段线性函数来进行分析，它使用一个矩阵乘法都为线性的理论模型。但实际上，由于浮点算法的存在，计算机上的矩阵乘法并非线性的。

Huang 等人[22]研究了图像分类器的安全性，例如划痕、相机角度或照明条件的变化对分类结果造成的影响，以及在原始图像的较小邻域内的分类不变性。他们为前馈多层神经网络开发了一种新的自动验证框架，利用 SMT 将输入样本周围的无限区域离散到一组点，然后在这个区域内进行有限穷举搜索并逐层传播分析，以验证神经网络的局部鲁棒性。Ehlers 等人[21]考虑了整个神经网络的全局线性逼近，并使用整数算法来提升 SMT 线性逼近范围的精确性，减少了 SMT 求解器的调用次数。

Narodytska 等人[49]研究了二值神经网络（Binarized Neural Networks）[50]的鲁棒性验证问题，将二值神经网络编码为布尔公式，然后利用布尔可满足性（Boolean Satisfiability，SAT）问题的典型求解方法进行求解，或是基于反例引导（counterexample-guided）搜索的思想，更有效地利用这些编码的结构来求解最终的 SAT 公式[51]。这种编码是深度神经网络的第一个精确布尔表示，不依

赖于网络结构的任何近似值，这意味着这种编码使我们能够通过研究 SAT 域中的相似属性来研究 BNN 的属性，并且将这些属性从 SAT 映射回神经网络域是精确的。但是这种方法只适用于二值神经网络，难以扩展到其他类型的神经网络。

然而，这些基于 SMT 的方法具有很高的计算复杂度，仅对非常小的网络才有意义。此外，由于 SMT 难以处理非线性运算，因此基于 SMT 的方法通常只适用于激活函数为分段线性的神经网络。

3.2 基于混合整数线性规划的方法

Cheng 等人[52]将计算模型鲁棒性边界的问题形式化为混合整数规划（Mixed Integer Programming, MIP）问题，并且设计了一系列启发式算法进行网络函数的编码，以大大减少 MIP 求解器的运行时间。具体地，他们利用基于分支定界算法的 MIP 求解器处理含有整数变量的非线性函数的编码，利用著名的 big-M 编码策略[53]的变体将许多非线性表达式线性化。此外，他们还定义了一个数据流分析[54]，该数据流分析用于生成相对较小的 big-M 作为加速 MIP 解决的基础。

Lomuscio 等人[55]研究了 ReLU 前馈神经网络的可达性问题，并将其形式化为线性规划问题求解。具体地，他们提出的方法能够检查一个特定的输出，比如一个漏洞，是否能够由一个给定的神经网络产生，而对抗样本则可以被视为一种特殊的可达性情况，在这种情况下输入集相对于特定输入是受限的，因此他们提出的公式具有一定的通用性。特别地，他们对线性规划中的浮点运算处理方式进行优化，并采用了一种较为高效的线性规划求解方法，相比于 Reluplex[15]中作为对比的线性规划方法效率大大提升。例如，Reluplex 只能分析最多 300 个 ReLU 约束条件，而他们的方法能够轻松处理 500 多个 ReLU 约束条件。Xiang 等人[56]研究了多层感知机（Multi-Layer Perception, MLP）的可达性问题和鲁棒验证问题，基于仿真验证思想[57]，利用大量仿真结果中的信息估计 MLP 的输出可达集并进行鲁棒性验证。具体地，他们引入了称为最大灵敏度的概念，对于激活函数为单调函数的一类多层感知机，将最大灵敏度的计算公式形式化为一组凸优化问题，然后通过解决凸优化问题来计算最大灵敏度。然后，利用结果获得最大灵敏度，通过检查有限数量的 MLP 采样输入的最大灵敏度属性来执行 MLP 的可达集合估计。最后，基于输出可达集的估计结果开发了一个自动验证 MLP 的方法，并介绍了安全验证在具有两个关节的机器人手臂模型中的应用。

Bunel 等人[58]利用分支定界理论[59]将现有方法纳入到一个统一框架中，收集并公开了一个基准数据集（benchmark），其中包含现有方法所用到的测试用例以及新的 benchmark，并使用这一基准数据集对现有算法进行首次大规模实验比较。在此框架的基础上，他们发现了验证算法的可改进之处，特别是在计算边界的方式、考虑的分支类型以及指导分支的策略上。与先前的技术相比，这些改进实现了将近两个数量级的加速。Tjeng 等人[23]将分段线性神经网络的鲁棒性验证问题形式化为混合整数线性规划问题，通过精心设计的预求解方法和有效的剪枝算法显著地减少了搜索空间，相比于基于 SMT 的方法计算速度提升了若干个数量级，因此可以用于计算含有超 100,000 个神经元的神经网络的鲁棒边界。特别地，他们首次证明了 MNIST 分类器在有界ℓ_∞范数$\epsilon = 0.1$扰动下的精确对抗精度：4.38%的测试样本在这个扰动范围内至少存在一个对抗样本，而其余测试样本在这个扰动范围内是鲁棒的，即不存在对抗样本。

虽然这类方法在小规模网络上取得了不错的结果，但是将它们扩展到更大规模的网络仍然是一个极具挑战性的问题。此外，这种方法也只适用于分段线性神经网络，即神经网络中只包含最大池化层和 ReLU 激活函数这两种非线性形式。

4 近似方法

由于在$\ell_p - ball$假设空间内，对于激活函数为 ReLU 的神经网络，计算其精确的鲁棒性边界是一个 NP 完备（NP-Complete，NPC）问题[15]，因此大多数研究者通常利用近似方法计算模型鲁棒性边界的下界，下文提到模型鲁棒性边界时通常也指的是这个下界。此外，对抗攻击[12]可以得到模型鲁棒性边界的上界[24]。因此，精确的模型鲁棒性边界可以由上界和下界共同逼近。这类方法通常基于鲁棒优化思想，通过解决公式（1）的内层最大化问题来估计模型鲁棒性边界：

$$\theta = \underset{\theta}{\operatorname{argmin}} \max_{\tilde{x} \in D_k(x)} L(y, f_\theta(\tilde{x})) \quad (1)$$

其中，x代表正常样本，$\tilde{x}$代表对抗样本，$D_k(x)$代表

对抗样本可能存在的范围，y代表样本真实标签，f_θ代表以θ为参数的模型，L代表损失函数。图 3 梳理了典型模型鲁棒性近似分析方法的相关研究工作。

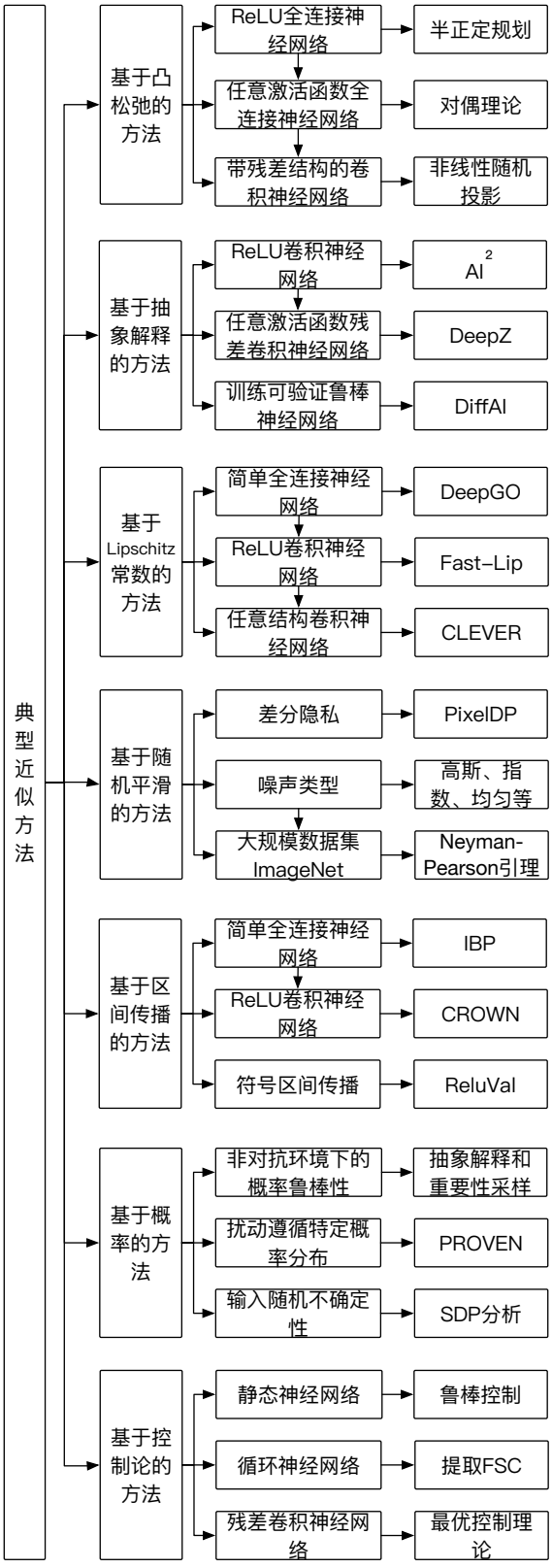


图 3：典型模型鲁棒性近似分析方法概念图。

4.1 基于凸松弛的方法

Wong 等人[26]采用凸多面体松弛方法来计算在范数有界扰动下的模型鲁棒边界，并将其形式化为一个线性规划问题求解。如图 4 所示，左图中黑色的点代表原始样本，正方形表示可扰动范围，即对抗样本可能存在的区域。经过神经网络的层层计算之后，正方形变成了非凸的多边形，因此他们的主要思想就是求多边形边框的凸近似，将非凸优化转换成凸优化问题。此外，他们证明了该线性规划的对偶问题可以用与反向传播网络相似的深层网络表示，从而提出了一种非常有效的优化方法，给出了有理论保证的鲁棒边界。但是，这种方法只讨论了 ReLU 这一种分段线性激活函数和最简单的前馈神经网络，未考虑其他类型尤其是非线性激活函数，也并未考虑带有卷积层或残差结构的复杂神经网络。并且，这种方法计算复杂度较高，在最坏情况下计算复杂度与神经元的个数成平方，因此也不适用于在 ImageNet 等大规模数据集上训练的大型神经网络。

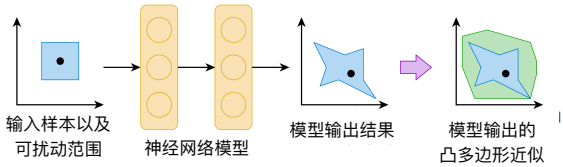


图 4：Wong&Kolter[26]采用的凸多面体松弛方法示意图。

为了克服上述挑战，Wong 等人[60]改进了上述方法，将其扩展到了带有卷积层或残差结构的复杂神经网络，并且神经网络的激活函数不再局限于 ReLU。具体地，他们采用了非线性随机投影（Random Projection）理论[61]，提出了一种基于共轭函数的、从对偶结构中派生对偶网络的方法，使得算法的计算复杂度降低至与网络的神经元个数成线性关系。与[26]提出的共轭形式不同的是，[60]不再假设网络由线性操作和激活函数组成，而是选择使用任意 k 函数序列，这简化了对典型神经网络架构中常见的最大池化（max pooling）等非线性激活函数的分析。其次，[60]中神经网络的隐藏层不仅依赖于前一层，还依赖于前一层的所有层，因此这种方法适用于具有任意跳跃连接（skip connection）的网络，例如残差（residual）网络和稠密（dense）网络。但是，这种方法仍然不适用于在 ImageNet 等大规模数据集上训练的大型神经网络。

Dvijotham 等人[27]基于优化理论和对偶（duality）理论的思想，将模型的鲁棒性验证问题形式化为一个无约束凸优化问题，然后通过解决这一无约束凸优化问题的拉格朗日松弛来获得可证明鲁棒边界。由于计算鲁棒边界的过程仅涉及解决一个无约束的凸优化问题，因此可以使用次梯度方法来求解。因此，对于任何对偶变量的选择，这种方法都可以获得一个有意义的鲁棒边界。虽然他们的方法适用于含有任意激活函数的神经网络，但是只适用于前馈神经网络，而不适用于带有卷积层或残差结构的复杂神经网络。

Raghunathan 等人[25]提出了一种基于半正定松弛（Semi-Definite Relaxation，SDR）的模型鲁棒性分析方法，即通过对原问题进行松弛，然后转化为最大割问题（Max-Cut Problem），最后用半正定规划（Semi-Definite Programming，SDP）[62]求解。他们不仅计算了对半径为ϵ的ℓ_∞球内所有输入（包括对抗输入）的鲁棒性边界，而且将鲁棒边界纳入模型训练的损失函数中，用于提升模型的鲁棒性。但是，他们只研究了具有二个隐藏层的前馈神经网络，而未考虑层数更多的前馈神经网络或具有卷积层的神经网络。随后，Raghunathan 等人[63]改进了先前的方法，同时提升了其分析精度与可扩展性。具体地，他们将基于半正定规划的方法扩展到具有任意层数的神经网络，并且考虑了中间层激活函数的相关性，因此精度相比于未考虑这种相关性的基于线性规划（Linear Programming，LP）的松弛方法[26,27,64]有了显著提升。此外，他们在理论上证明了对于具有随机权重的神经网络，LP 松弛和 SDP 松弛之间存在平方根维度的差距。Fazlyab 等人[65]也提出了一种基于 SDP 的鲁棒性分析框架，利用二次约束的形式化来抽象激活函数的各种属性，例如单调性、有界斜率、有界值和跨层重复，然后通过 SDP 来分析抽象网络的鲁棒性。对于深层网络来说，这种方法的计算复杂度比先前的方法降低了一个数量级，并且适用于任意激活函数。

Jordan 等人提出了 GeoCert[66]，将分段线性神经网络的输入空间划分为多面复合体（polyhedral complices），并将计算 ReLU 神经网络的逐点鲁棒性问题与计算具有固定中心的最大范数球的问题联系起来。通过不断优化输出的 point-wise 鲁棒边界，GeoCert 最终能够输出一个紧密的鲁棒边界。Salman 等人[67]提出了一个分层的凸松弛框架，统一了 DeepZ、DeepPoly、CROWN、Fast-Lin 等方法并揭示了它们之间的关系，并进一步表明在此框架内方法的性能在理论上受到限制，即一味的追求更精确的分层凸松弛方法并不能突破这个理论上限。在不同的模型、训练方法和数据集上，他们发现：（i）最佳的分层凸松弛仅略微改进了 Wong 和 Kolter[26]发现的模型鲁棒性边界，与 PGD 攻击提供的边界相比始终大 1.5 到 5 倍；（ii）就鲁棒误差的上限而言，最优的逐层凸松弛不会显著缩小 PGD 下限与 Wong 和 Kolter[26]提出的上限之间的差距。

4.2 基于抽象解释的方法

这一类方法使用抽象解释（Abstract Interpretation）[54]来验证具有分段线性激活的神经网络的鲁棒性，典型方法包括 AI^2[34]、DeepZ[35]、DiffAI[36]等。它们使用抽象域（Abstract Domain），即某些几何形状（例如 box，zonotope，polytope 等）的一组逻辑约束，来近似神经网络的输入层、隐藏层以及输出层，从而验证模型的鲁棒性。因此，广义上这种方法可以理解为在神经网络的前向传播中维持激活函数组合的松弛。

Pulina 等人[68]最早将抽象解释理论应用于模型鲁棒性分析，但是他们的方法只在仅有 6 个神经元的网络上试验成功。为了将抽象解释应用于分析规模更大、更复杂的神经网络，Gehr 等人提出了 AI^2[34]，首先定义了 zonotope 抽象域以捕获所有潜在的对抗输入，然后为 CNN 的每个非线性操作创建了抽象转换器（Abstract Transformer），利用抽象转换器在目标 CNN 中传播抽象域，最后利用输出抽象域的范围来验证模型的鲁棒性。然而，AI^2 只适用于激活函数为 ReLU 的分段线性网络，并且它对 ReLU 函数的近似精度和效率都较低。为了扩展到其他类型的神经网络，Singh 提出了 DeepZ[35]，其适用于激活函数为 ReLU、Tanh、Sigmoid 等类型的神经网络，并且能够处理 ResNet 等带有残差结构的模型。

为了进一步平衡精度和可扩展性，Mirman 等人提出了如图 5 所示的包含一系列新型抽象转换器的 DiffAI[36]并将其用于训练可验证鲁棒的模型，然后在后续的工作中[69]引入用于微调抽象精度和可扩展性的抽象层，以及用于描述将抽象损失与具体损失相结合的目标函数的领域专用语言（Domain Specific Language，DSL），然后将其用于 ResNet-34

和 DenseNet-100 等大规模网络的鲁棒训练。这些方法都是采用 zonotope 作为抽象域，因此在精度方面存在天然缺陷。

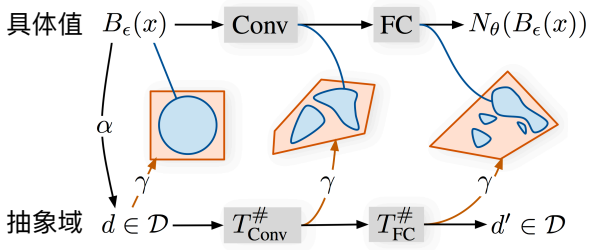


图 5：基于抽象解释的模型鲁棒性分析方法 DiffAI[36]应用于具有卷积和完全连接层的神经网络的工作流程示意图。

为了克服这种缺陷，Singh 等人提出了 DeepPoly[37]，将抽象域从之前的 zonotope 换成了新型的 polyhedra，进一步提升了这类方法的精度与可扩展性。此外，Singh 等人提出的 RefineAI[38]，将抽象解释和混合整数线性规划结合起来，进一步提升了这类方法的精度。最近，Singh 等人提出 k-ReLU[39]，利用抽象解释近似神经网络时统一考虑多个 ReLU 激活函数的输出而不是分开单独计算，并利用它们之间的相关性进一步提升相关方法的精度。

4.3 基于Lipschitz常数的方法

Szegedy 等人[12]计算了每一层的全局 Lipschitz 常数，并使用其乘积来解释神经网络中的鲁棒性问题。具体地，令x表示输入，W_k为模型第 k 层的参数，$\phi(x)$为模型的输出，ϕ_k为模型第 k 层的运算，即$\phi(x) = \phi_K(\phi_{K-1}(\dots\phi_1(x;W_1);W_2)\dots;W_K)$，那么$\phi(x)$的不稳定性可以用每一层的 Lipschitz 常数来表示，即 $\forall x, r, \|\phi_k(x;W_k) - \phi_k(x+r;W_k)\| \le L_k\|r\|$，其中$L_k > 0, k = 1,2,\dots,K$。那么，神经网络的最终输出就满足$\|\phi(x) - \phi(x+r)\| \le L\|r\|$，其中$L = \prod_{k=1}^{K} L_k$。然而，全局 Lipschitz 常数给出的边界通常十分宽松。

为了得到更精确的边界，Hein & Andriushchenko[24]使用局部 Lipschitz 连续条件给出了模型鲁棒空间的边界：$\epsilon = \max_{R>0} \min\{\min_{j\neq c} \frac{f_c(x)-f_j(x)}{\max_{y\in B_p(x,R)} \|\nabla f_c(y) - \nabla f_j(y)\|_q}, R\}$，其中$\frac{1}{p} + \frac{1}{q} = 1, B_p(x,R) = \{y \in \mathbb{R}^d \mid \|x-y\|_p \le R\}$。基于此，他们推导了只有一个隐藏层和 softplus 激活函数的多层感知机（Multi-Layer Perception, MLP）的闭合形式边界。然而，对于具有多个隐藏层的神经网络，这种方法很难得出闭合形式的边界。基于类似的思想，Croce 等人[70]将[71]中针对ℓ_2范数和ℓ_∞范数的方法扩展到了ℓ_p范数，其中$p \in [1,\infty]$，证明了对决策边界的ℓ_1和ℓ_∞范数的保证足以在所有ℓ_p范数的鲁棒性上得出有意义的证明，并且这种保证与输入空间的大小无关。

Ruan 等人提出了 DeepGO[72]，将神经网络的鲁棒空间分析转化为可达性问题，并利用自适应优化（Adaptive Optimization）方法来解决这个可达性问题。由于网络和函数是 Lipschitz 连续的，即上下限之间的所有值都是可达的，因此对于给定的输入数据集，神经网络的鲁棒空间分析就可以转化为计算其输出值的 Lipschitz 连续函数的下限和上限。具体地，他们的方法不是直接将模型的参数以及激活函数转化为一系列线性约束条件，而是需要计算这个模型的 Lipschitz 常数。他们证明了不能直接转化为线性约束条件的隐藏层是 Lipschitz 连续的，并且通过分析激活函数和相应参数能够计算一个精确的 Lipschitz 常数。但是，他们只研究了最简单的前馈神经网络模型。

为了得到多层复杂神经网络模型的鲁棒边界，Weng 等人[28]提出了 Fast-Lin 和 Fast-Lip，其中 Fast-Lip 是基于一系列特有的边界自变量计算神经网络模型局部 Lipschitz 常数得到模型的鲁棒边界，并且可以在数十秒内完成对一个具有 10,000 以上神经元的 7 层神经网络的鲁棒边界计算。在小型网络上，这种方法比 Reluplex 快了 10,000 倍以上；在大型网络上，这种方法比基于线性规划的方法快了 33~14,000 倍。然而，这种方法只适用于具有 ReLU 激活函数的神经网络。为了适用于不同结构的神经网络模型，Weng 等人[31]基于极值理论（Extreme Value Theory，EVT），提出一种新型模型鲁棒性度量标准 CLEVER（Cross-Lipschitz Extreme Value for nEtwork Ro-bustness），提供了将鲁棒边界分析转换为局部 Lipschitz 常数估计问题的理论依据。随后，Weng 等人[73]又提出了对 CLEVER 的两个扩展：(1)将极值理论应用于新的形式鲁棒性保证，为可二次区分的分类函数提供了新的形式上的鲁棒性保证，估计的鲁棒性称为二阶 CLEVER 得分；(2)讨论了如何使用带有后向通过可微分逼近（Backward Pass Differentiable Approximation，BPDA）的 CLEVER 处理常见的梯度掩模（gradient masking）防御技术。

借助 BPDA，CLEVER 可以评估更广泛类别的神经网络的内在鲁棒性，例如具有不可微分输入转换的网络。

Latorre 等人[74]提出了一种基于多项式优化问题（Polynomial Optimization Problem, POP）[75]松弛的计算神经网络 Lipschitz 常数上限的一种通用方法 LiPopt，这种方法利用多项式不等式描述单位球（unit ball），因此同时涵盖了ℓ_2范数和ℓ_∞范数。基于 Weisser 等人提出的定理[76]，作者利用神经网络架构的稀疏连通性来推导一系列线性规划问题，并根据神经元的数量、网络的深度和稀疏性对此类 LP 问题进行渐进分析。

4.4 基于随机平滑的方法

虽然研究者提出了许多模型鲁棒边界分析方法，但是这些方法很少能够扩展到在 ImageNet 等极具挑战性的数据集上训练的大型神经网络，并且适用的模型种类也非常有限。为了解决这一问题，Lecuyer 等人[32]提出了 PixelDP，利用差分隐私和模型鲁棒性之间的联系首次提出了“随机平滑（Randomized Smoothing）”这一概念，提供了对 top-1 预测进行随机平滑处理的鲁棒性保证。具体地，以测试阶段为例，对于给定的样本x，传统的神经网络分类模型将单次预测得到的 argmax 类别作为最终预测结果给出，即$f(x) = argmax\, f_j(x), j = 1,2,\dots,K$，而随机平滑方法则是利用蒙特卡洛（Monte Carlo）估计方法，通过在样本x周围采样n个样本$x_1, x_2, \dots, x_n$（相当于给x添加随机噪声）得到它们 argmax 类别的期望值并将其作为最终预测结果，即$\hat{f}(x) = \frac{1}{n}\sum_n f(x_n)$。Li 等人[77]在高斯随机噪声的假设下，得出了模型鲁棒边界的表达形式：$L = \sup_{\alpha>1}\left(-\frac{2\sigma^2}{\alpha}\log\left(1 - p_{(1)} - p_{(2)} + 2\left(\frac{1}{2}\left(p_{(1)}^{1-\alpha} + p_{(2)}^{1-\alpha}\right)\right)^{\frac{1}{1-\alpha}}\right)\right)^{1/2}$，其中高斯噪声服从分布$N(0,\sigma^2)$，$p_{(1)}$和$p_{(2)}$分别是期望值第一和第二大的两个类别的概率值。但是，上述两种方法计算得到的鲁棒性边界十分宽松，并且没有实验证明它们适用于在 ImageNet 上训练的神经网络模型。

为了得到更精确的鲁棒性边界，Cohen 等人[33]利用 Neyman-Pearson 引理获得了高斯随机噪声假设下的随机平滑的可验证$\ell_p - ball$半径范围，即$R = \frac{\sigma}{2}(\Phi^{-1}\left(\underline{p_1}\right) - \Phi^{-1}(\overline{p_2}))$，其中$\Phi^{-1}$是标准高斯函数累计分布函数（Cumulative Distribution Function，CDF）的逆函数，$\mathbb{P}(f(x+\epsilon) = c_1) \geq \underline{p_1} \geq \overline{p_2} \geq \max_{c \neq c_1}\mathbb{P}(f(x+\epsilon) = c)$。这种方法得到的鲁棒性边界的精确度大幅度优于先前的方法[29,60,78]，并且当$\ell_2 < 127/255$时，该方法在 ImageNet 数据集上实现了 49%的可验证准确率。Pinot 等人[79]进一步从理论上证明了随机平滑方法的有效性，并且给出了当噪声服从指数分布时经随机平滑处理的模型鲁棒性边界。Lee 等人[80]对扰动的分布类型进行了扩展，从将[33]中的ℓ_2范数连续空间假设扩展到ℓ_0范数离散空间假设。此外，研究者还对噪声服从均匀分布[81]、多项式分布[82]假设的模型鲁棒性边界做了初步的探讨。

Salman 等人[83]为随机平滑分类器设计了一种自适应攻击，并在对抗训练环境中使用这种攻击来增强平滑分类器的可证明的鲁棒性。Dvijotham 等人[84]将这种思想扩展到了任意平滑方法，利用 f 散度证明了平滑分类器的鲁棒性。Salman 等人[85]针对预训练模型提出了一种黑盒随机平滑的方法，通过在现有的图像分类模型之前添加降噪器并使用随机平滑得到新分类模型，在不修改预训练分类模型的情况下保证了对对抗样本具有ℓ_p鲁棒性，可用于保护 Azure、Google、AWS 和 ClarifAI 等图像分类 API。Jia 等人[86]考虑了在实际中应用更为广泛的 top-k 预测鲁棒性。Wang 等人[87]和 Maurice 等人[88]基于随机平滑的思想研究了模型对后门攻击（backdoor attack）的鲁棒性。Mohapatra 等人[89]首次指出了当前随机平滑方法的副作用：1）决策边界随着随机平滑预测规则的采用而缩小；2）增强噪音并不一定能解决边界收缩的问题，甚至会产生其他问题。

4.5 基于区间边界传播的方法

Gowal 等人[40]首次提出了基于区间边界传播（Interval Bound Propagation，IBP）思想[90]的可验证鲁棒神经网络训练方法，当输入数据在“∞范数限制的球内”被扰动时，IBP 定义的损失函数以最小化任何一对对数之间的最大差的上界。与更复杂的方法[26,36,64]相比，IBP 的速度非常快，其计算成本可与网络的两次前向传递相媲美，这使得这种方法具有很高的可扩展性。例如，Huang 等人[43]证明 IBP 的思想还可用于分析自然语言处理模型的鲁棒性，Chiang 等人[91]将 IBP 用于分析针对补丁（patch）型

对抗攻击的模型鲁棒性。虽然这种方法在某些任务上明显优于基于线性松弛的方法，但是由于边界要宽松得多，因此存在稳定性问题。为了克服这一缺陷，Zhang 等人通过组合前向传播的 IBP 边界[40]和后向传播基于紧线性松弛的边界 CROWN[29]，提出了一种新的具有鲁棒性保证的对抗训练方法 CROWN-IBP[92]，计算效率非常高，并且在训练可验证的鲁棒神经网络方面始终优于 IBP 方法。

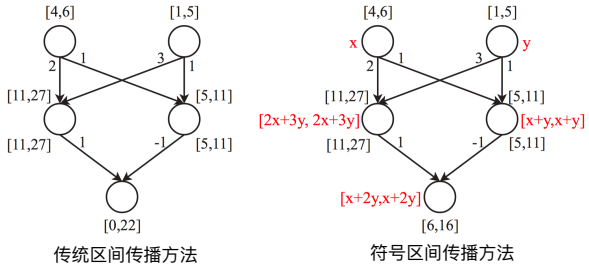


图 6：传统区间传播方法与符号区间传播方法[93]对比。

最近，Wang 等人提出了 ReluVal[93]，利用符号区间传播的方法来验证模型的鲁棒性。图 6 展示了传统区间传播方法与符号区间传播方法之间的差别。从图 6 中我们可以看出，传统区间传播方法的输出间隔非常宽松，因为它忽略了输入变量的相互依赖关系，而使用符号区间分析则可以跟踪变量之间的依赖关系，因此计算得到的边界会更加精确。然而，这种方法的计算复杂度较高，因此只适用于只具有几个隐藏层的小规模简单神经网络，难以扩展到实际中的复杂大规模神经网络。在后续的工作中[94]，Wang 等人进一步优化了对 ReLU 非线性激活函数的松弛方法，得到了更紧密的鲁棒性边界。

4.6 基于概率的方法

与最坏情况下的模型鲁棒性验证方法相比，在概率环境中研究神经网络验证的工作相对较少[95-98]。Mangal 等人[95]认为大多数现有的鲁棒性定义都集中在最不利的情况下，这种鲁棒性的概念太强了，不太可能被实用的神经网络所满足和验证。因此，他们根据输入空间在实际中的潜在非对抗性概率分布来生成神经网络的输入，并且只研究非对抗环境下神经网络对可能在现实世界中生成的对抗输入的鲁棒性。他们提出了一种新的鲁棒性概念——概率鲁棒性，它要求神经网络对输入分布的鲁棒性至少应为$1-\varepsilon$概率。基于此，他们提出了一种基于抽象解释和重要性采样（importance sampling）的用于验证神经网络是否具有概率鲁棒性的算法。具体地，他们首先用抽象解释来近似神经网络的行为，并计算违反鲁棒性的输入区域的过度近似，然后使用重要性采样来抵消这种过度近似的影响，并计算出神经网络违反鲁棒性的概率的准确估计。

基于类似的思想，Weng 等人提出了 PROVEN[96]，研究了在扰动遵循特定的概率分布时模型的概率鲁棒性，从统计意义上提供了对任意的有界ℓ_p扰动，模型的 top-1 预测结果不会改变的概率保证。此外，他们表明有可能基于当前最新的神经网络鲁棒性验证框架来得出封闭形式的概率证明，因此当从现有方法（例如 Fast-Lin[28]，CROWN[28]和 CNN-Cert[28]）获得最坏情况下的经过认证的鲁棒性边界时，PROVEN 提供的概率保证能够自然而然地、以几乎没有其他计算代价的方式产生。在小型和大型 MNIST 和 CIFAR 神经网络模型上进行的实验表明，与 CROWN 提供的最坏情况的鲁棒性证书相比，PROVEN 可以将鲁棒性证书的性能提高多达 75%左右，至少达到 99.99%。

Fazlyab 等人[97]假设输入不确定性是随机且无限的，这种随机不确定性可能是由于数据量化、输入预处理或环境背景噪声等因素导致的[96]，并在这种设定条件下研究了两个模型鲁棒性相关的问题：（i）概率安全验证，即给定神经网络输出空间中的安全区域，估计当神经网络的输入受到均值和协方差已知的随机噪声干扰时，其输出将处于安全区域的可能性；（ii）置信度椭球估计，即给定神经网络输入的置信度椭球，估计输出的置信度椭球。由于存在非线性激活函数，因此很难精确地解决这两个问题。为了简化分析，他们的主要思想是通过将非线性激活函数强加于它们的输入-输出对的抽象和二次约束的组合来抽象它们，然后使用 SDP 来分析满足原始网络安全性的抽象网络的安全性，为神经网络的输出提供统计保证。

4.7 基于控制论的方法

Wang 等人[99]研究了如何将静态神经网络验证工具与鲁棒控制理论相结合，以在控制回路中对神经网络鲁棒性进行验证。具体地，给出一个充分的条件和一种算法，他们确保当对抗扰动为∞范数有界时，闭环状态和控制约束得到满足。他们的方法基于闭环动力系统的一个正不变集，因此不需要神经网络满足可微性或连续性。在车杆控制（cart pole control）问题上，这种方法证明的神经网络鲁棒性与传统的基于 Lipschitz 常数的方法相比要好 5 倍。

Carr 等人[100]通过集成形式化方法和机器学习技术，提出了一种从 RNN 中自动提取有限状态控制器（finite-state controller，FSC）的方法，当与有限状态系统模型组成时，该状态控制器适用于现有的模型鲁棒性形式化验证工具。具体地，他们对 quantized bottleneck insertion 技术进行了迭代修改，以创建 FSC 作为具有内存的随机策略。对于 FSC 无法满足规范的情况，他们利用诊断信息来调整提取的 FSC 中的内存量或重新训练 RNN。

Wang 等人[101]将运输方程的最优控制理论与 ResNets 的训练和测试实践相统一，并基于这种统一的观点提出了一种简单而有效的 ResNets 集成算法，以提高在干净图像和对抗图像上经过严格训练的模型的准确性。所提出的算法包括两个部分：首先，通过将方差指定的高斯噪声注入每个残差映射的输出来修改基础的 ResNet；其次，对多个协同训练的经过修改的 ResNet 的生产进行平均，以获得最终预测。

综上，典型的模型鲁棒性分析方法优缺点总结如表 2 所示。

表 2 典型的模型鲁棒性分析方法优缺点总结

方法	优点	缺点
基于可满足性模理论的方法	能够精确验证神经网络的鲁棒性	效率非常低，计算成本太高，只适用于小规模分段线性神经网络，无法扩展到实际中的复杂大规模神经网络
基于混合整数线性规划的方法	能够精确验证神经网络的鲁棒性；相比于基于 SMT 的方法，效率有所提高，能够适用于规模稍大的神经网络	仍然只能处理分段线性神经网络，无法处理含有非线性激活函数的神经网络；适用的神经网络规模仍然不足以支撑实际应用
基于凸松弛的方法	相比于精确方法，效率有所提高，能够适用于规模稍大的神经网络和含有非线性激活函数的神经网络	性能在理论上受到限制，即一味的追求更精确的分层凸松弛方法并不能突破这个理论上限；不适用于在 ImageNet 等大规模数据集上训练的大型神经网络
基于抽象解释的方法	考虑了变量之间的相关性，计算精度相对较高；能够处理非线性激活函数和带有残差结构的模型	精度会随着神经网络层数的增加而下降，而实验中的神经网络规模层数普遍不超过 10 层，因此应用于实际中几十层神经网络时的精度可能会大大下降
基于 Lipschitz 常数的方法	适用的神经网络类别更广泛，例如具有不可微分输入转换的网络	计算得到的鲁棒边界较为宽松
基于随机平滑的方法	依靠在噪声假设下传统模型进行良好决策的能力，将鲁棒分类问题扩展为经典监督学习问题，适用于任意结构的神经网络	决策边界随着随机平滑预测规则的采用而缩小；增强扰动并不一定能解决边界收缩的问题，甚至会产生其他问题
基于区间边界传播的方法	计算速度快，具有很高的可扩展性	虽然这种方法在某些任务上明显优于基于线性松弛的方法，但是由于边界要宽松得多，因此存在稳定性问题
基于概率的方法	具有更宽松的鲁棒性定义，更有可能被实用的神经网络所满足和验证	对噪声分布的假设通常在实际中较难满足
基于控制论的方法	与传统的基于 Lipschitz 常数的方法相比精度有较大提升	适用的范数扰动有限

5 未来研究方向

本文介绍了模型鲁棒性分析问题的背景与挑战，探讨了相关定义，进而对目前主流的模型鲁棒性方法与性能做了介绍。从目前已有的相关方法来看，我们认为今后对模型鲁棒性分析方法的研究，将主要围绕以下几个方向展开：

（1）进一步拓展对抗扰动的类型。从攻击者添加扰动的类型来看，现存的大多数模型鲁棒性方法都是针对在像素点上添加扰动的对抗攻击进行的鲁棒性分析，然而在实际中，对抗性图像有可能经过旋转、缩放等几何变换，而现存大多数方法无法扩展到此类变换。虽然 Balunovic 等人提出的 DeepG[102]初步尝试了将抽象解释的思想用于分析几何变换对抗攻击的模型鲁棒性空间，但是这个方向仍然值得更多深入研究，进一步提升精度和可扩

展性。

（2）不同鲁棒性类型之间的平衡。输入样本x的局部鲁棒性（即神经网络应为以x为中心的半径为ϵ的球中的所有输入产生相同的预测结果）依赖于在输入空间上定义的合适的距离度量标准，在实际中，对于在非恶意环境中运行的神经网络而言，这可能是太过苛刻的要求。同时，由于仅针对特定输入定义了局部鲁棒性，而对于未考虑的输入不提供保证，因此局部鲁棒性也具有固有的限制性。全局鲁棒性则通过进一步要求输入空间中的所有输入都满足局部鲁棒性来解决这个问题。除了在计算上难以控制之外，全局鲁棒性仍然太强而无法实际使用。因此，在实际中如何更好地平衡局部鲁棒性与全局鲁棒性，仍然是一个亟待解决的挑战。

（3）进一步提升模型鲁棒性验证方法。从实证结果来看，大多数基于经验的防御方法非常容易被更强的攻击所攻破，而其他鲁棒性分析方法在很大程度上取决于神经网络模型的体系结构，例如激活函数的种类或残差连接的存在。相比之下，随机平滑不对神经网络的体系结构做任何假设，而仅依靠在噪声假设下传统模型进行良好决策的能力，从而将鲁棒分类问题扩展为经典监督学习问题，可用于社区检测[103]等任务。因此，基于随机平滑的鲁棒性分析方法可能是研究模型鲁棒空间的最有前途的方向之一。此外，由于基于概率的方法具有更宽松的鲁棒性定义，更有可能被实用的神经网络所满足和验证，因此在合适的扰动分布假设下也是较有前景的方向之一。

（4）研究可证明鲁棒模型训练方法。此外，如何训练对对抗性扰动具有可证明鲁棒的神经网络以及如何训练更容易验证鲁棒性的神经网络，也是未来的研究方向之一。目前研究者在这个方向进行的初步探索包括利用正则化技术将模型的形式化鲁棒边界与模型的目标函数结合起来[104]、经验性对抗风险最小化（Empirical Adversarial Risk Minimization，EARM）[36,105]、随机自集成[106]、剪枝[82,107]以及改善神经网络的稀疏性[108]。但是现存技术主要集中于图像领域，难以扩展到恶意软件等安全攸关型应用，并且仍然存在精度以及可扩展性上的不足，需要进一步的深入研究。

6 结束语

随着深度学习研究进一步发展和深度学习技术在实际场景中的广泛应用，深度学习模型的鲁棒性研究成为了一个新生而又有前景的研究领域，吸引了一大批来自于学术界和工业界的学者的广泛兴趣和深入研究，并且取得了许多瞩目的研究成果。然而，到目前为止，深度学习模型的鲁棒性研究还处于初级阶段，依然存在许多关键的科学问题尚待解决。为了重新审视深度学习模型的鲁棒性研究现状，理清现有研究成果的优势与不足，明确未来研究方向，本文系统地研究了深度学习模型的鲁棒性问题，回顾了大量的极具影响力的研究成果并对相关研究进行了科学的分类、总结和分析。同时，本文指出了深度学习模型的鲁棒性研究当前面临的挑战，探讨了未来可行的研究方向，旨在推动深度学习模型的鲁棒性分析研究的进一步发展。

参 考 文 献

[1] Wang T, Gao H, Qiu J. A combined adaptive neural network and nonlinear model predictive control for multirate networked industrial process control. IEEE Transactions on Neural Networks and Learning Systems, 2015, 27(2): 416-425.

[2] Schmidhuber J. Deep learning in neural networks: An overview. Neural Networks, 2015, 61: 85-117.

[3] Silver D, Huang A, Maddison C J, et al. Mastering the game of Go with deep neural networks and tree search. Nature, 2016, 529(7587): 484.

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

[5] Madry A, Makelov A, Schmidt L, et al. Towards deep learning models resistant to adversarial attacks. arXiv preprint arXiv:1706.06083, 2017.

[6] Papernot N, Mcdaniel P, Wu X, et al. Distillation as a defense to adversarial perturbations against deep neural networks//Proceedings of the 2016 IEEE Symposium on Security and Privacy, 2016: 582-597.

[7] Goodfellow I J, Shlens J, Szegedy C. Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572, 2014.

[8] Meng D, Chen H. Magnet: a two-pronged defense against adversarial examples//Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security, 2017: 135-147.

[9] Tramèr F, Kurakin A, Papernot N, et al. Ensemble adversarial training: Attacks and defenses. arXiv preprint arXiv:1705.07204, 2017.

[10] Das N, Shanbhogue M, Chen S-T, et al. Shield: Fast, practical defense and vaccination for deep learning using jpeg compression//Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2018: 196-204.

[11] Liao F, Liang M, Dong Y, et al. Defense against adversarial attacks using

high-level representation guided denoiser//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018: 1778-1787.

[12] Szegedy C, Zaremba W, Sutskever I, et al. Intriguing properties of neural networks//Proceedings of the International Conference on Learning Representations, 2014.

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

[14] Schott L, Rauber J, Bethge M, et al. Towards the First Adversarially Robust Neural Network Model on MNIST//Proceedings of the International Conference on Learning Representations, 2019: 1-16.

[15] 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, 2017: 97-117.

[16] Bastani O, Ioannou Y, Lampropoulos L, et al. Measuring neural net robustness with constraints//Proceedings of the Advances in Neural Information Processing Systems, 2016: 2613-2621.

[17] Pascanu R, Montufar G, Bengio Y. On the number of response regions of deep feed forward networks with piece-wise linear activations. arXiv preprint arXiv:1312.6098, 2013.

[18] Montufar G F, Pascanu R, Cho K, et al. On the number of linear regions of deep neural networks//Proceedings of the Advances in Neural Information Processing Systems, 2014: 2924-2932.

[19] Eldan R. A polynomial number of random points does not determine the volume of a convex body. Discrete & Computational Geometry, 2011, 46(1): 29-47.

[20] Zakrzewski R R. Verification of a trained neural network accuracy //Proceedings of the International Joint Conference on Neural Networks, 2001: 1657-1662.

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

[22] Huang X, Kwiatkowska M, Wang S, et al. Safety verification of deep neural networks//Proceedings of the International Conference on Computer Aided Verification, 2017: 3-29.

[23] Tjeng V, Xiao K, Tedrake R. Evaluating robustness of neural networks with mixed integer programming. arXiv preprint arXiv:1711.07356, 2017.

[24] Hein M, Andriushchenko M. Formal guarantees on the robustness of a classifier against adversarial manipulation//Proceedings of the Advances in Neural Information Processing Systems, 2017: 2266-2276.

[25] Raghunathan A, Steinhardt J, Liang P. Certified defenses against adversarial examples. arXiv preprint arXiv:1801.09344, 2018.

[26] Wong E, Kolter Z. Provable Defenses against Adversarial Examples via the Convex Outer Adversarial Polytope//Proceedings of the International Conference on Machine Learning, 2018: 5286-5295.

[27] Dvijotham K, Stanforth R, Gowal S, et al. A Dual Approach to Scalable Verification of Deep Networks//Proceedings of the Uncertainty in Artificial Intelligence, 2018, 1(2): 3.

[28] Weng L, Zhang H, Chen H, et al. Towards Fast Computation of Certified Robustness for ReLU Networks//Proceedings of the International Conference on Machine Learning, 2018: 5276-5285.

[29] Zhang H, Weng T-W, Chen P-Y, et al. Efficient neural network robustness certification with general activation functions//Proceedings of the Advances in Neural Information Processing Systems, 2018: 4939-4948.

[30] Boopathy A, Weng T-W, Chen P-Y, et al. CNN-Cert: An efficient framework for certifying robustness of convolutional neural networks//Proceedings of the AAAI Conference on Artificial Intelligence, 2019: 3240-3247.

[31] Weng T-W, Zhang H, Chen P-Y, et al. Evaluating the robustness of neural networks: An extreme value theory approach. arXiv preprint arXiv:1801.10578, 2018.

[32] Lecuyer M, Atlidakis V, Geambasu R, et al. Certified robustness to adversarial examples with differential privacy//Proceedings of the 2019 IEEE Symposium on Security and Privacy, 2019: 656-672.

[33] Cohen J, Rosenfeld E, Kolter Z. Certified Adversarial Robustness via Randomized Smoothing//Proceedings of the International Conference on Machine Learning, 2019: 1310-1320.

[34] Gehr T, Mirman M, Drachsler-Cohen D, et al. Ai2: Safety and robustness certification of neural networks with abstract interpretation//Proceedings of the 2018 IEEE Symposium on Security and Privacy, 2018: 3-18.

[35] Singh G, Gehr T, Mirman M, et al. Fast and effective robustness certification//Proceedings of the Advances in Neural Information Processing Systems, 2018: 10802-10813.

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

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

[38] Singh G, Gehr T, Püschel M, et al. Boosting Robustness Certification of Neural Networks//Proceedings of the International Conference on Learning Representations, 2019.

[39] Singh G, Ganvir R, Püschel M, et al. Beyond the Single Neuron Convex Barrier for Neural Network Certification//Proceedings of the Advances in Neural Information Processing Systems, 2019: 15072-15083.

[40] Gowal S, Dvijotham K D, Stanforth R, et al. Scalable Verified Training for Provably Robust Image Classification//Proceedings of the IEEE International Conference on Computer Vision, 2019: 4842-4851.

[41] Chen H, Zhang H, Si S, et al. Robustness verification of tree-based models//Proceedings of the Advances in Neural Information Processing Systems, 2019: 12317-12328.

[42] Wang Y, Jha S, Chaudhuri K. Analyzing the Robustness of Nearest Neighbors to Adversarial Examples//Proceedings of the International Conference on Machine Learning, 2018: 5133-5142.

[43] Huang P-S, Stanforth R, Welbl J, et al. Achieving verified robustness to symbol substitutions via interval bound propagation. arXiv preprint arXiv:1909.01492, 2019.

[44] Ko C-Y, Lyu Z, Weng L, et al. POPQORN: Quantifying Robustness of Recurrent Neural Networks//Proceedings of the International Conference on Machine Learning, 2019: 3468-3477.

[45] Jia R, Raghunathan A, Göksel K, et al. Certified Robustness to Adversarial Word Substitutions//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, 2019: 4120-4133.

[46] Shi Z, Zhang H, Chang K-W, et al. Robustness verification for transformers. arXiv preprint arXiv:2002.06622, 2020.

[47] Zügner D, Günnemann S. Certifiable robustness and robust training for graph convolutional networks//Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2019: 246-256.

[48] Bojchevski A, Günnemann S. Certifiable Robustness to Graph Perturbations//Proceedings of the Advances in Neural Information Processing Systems, 2019: 8317-8328.

[49] Narodytska N, Kasiviswanathan S, Ryzhyk L, et al. Verifying properties of binarized deep neural networks//Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, 2018.

[50] Hubara I, Courbariaux M, Soudry D, et al. Binarized neural networks//Proceedings of the Advances in Neural Information Processing Systems, 2016: 4107-4115.

[51] Clarke E, Grumberg O, Jha S, et al. Counterexample-guided abstraction refinement for symbolic model checking. Journal of the ACM, 2003, 50(5): 752-794.

[52] Cheng C-H, Nührenberg G, Ruess H. Maximum resilience of artificial neural networks//Proceedings of the International Symposium on Automated Technology for Verification and Analysis, 2017: 251-268.

[53] Grossmann I E. Review of nonlinear mixed-integer and disjunctive programming techniques. Optimization and engineering, 2002, 3(3): 227-252.

[54] Cousot P, Cousot R. Abstract interpretation: a unified lattice model for static analysis of programs by construction or approximation of fixpoints//Proceedings of the 4th ACM SIGACT-SIGPLAN symposium on Principles of programming languages, 1977: 238-252.

[55] Lomuscio A, Maganti L. An approach to reachability analysis for feed-forward relu neural networks. arXiv preprint arXiv:1706.07351, 2017.

[56] Xiang W, 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.

[57] Duggirala P S, Mitra S, Viswanathan M, et al. C2E2: A verification tool for stateflow models//Proceedings of the International Conference on Tools and Algorithms for the Construction and Analysis of Systems, 2015: 68-82.

[58] Bunel R R, Turkaslan I, Torr P, et al. A unified view of piecewise linear neural network verification//Proceedings of the Advances in Neural Information Processing Systems, 2018: 4790-4799.

[59] Lawler E L, Wood D E. Branch-and-bound methods: A survey. Operations research, 1966, 14(4): 699-719.

[60] Wong E, Schmidt F, Metzen J H, et al. Scaling provable adversarial defenses//Proceedings of the Advances in Neural Information Processing Systems, 2018: 8400-8409.

[61] Vempala S S. The random projection method. 65. American Mathematical Soc., 2005.

[62] Goemans M X, Williamson D P. Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming. Journal of the ACM, 1995, 42(6): 1115-1145.

[63] Raghunathan A, Steinhardt J, Liang P S. Semidefinite relaxations for certifying robustness to adversarial examples//Proceedings of the Advances in Neural Information Processing Systems, 2018: 10877-10887.

[64] Dvijotham K, Gowal S, Stanforth R, et al. Training verified learners with learned verifiers. arXiv preprint arXiv:1805.10265, 2018.

[65] Fazlyab M, Morari M, Pappas G J. Safety verification and robustness analysis of neural networks via quadratic constraints and semidefinite programming. arXiv preprint arXiv:1903.01287, 2019.

[66] Jordan M, Lewis J, Dimakis A G. Provable certificates for adversarial examples: Fitting a ball in the union of polytopes//Proceedings of the Advances in Neural Information Processing Systems, 2019: 14059-14069.

[67] Salman H, Yang G, Zhang H, et al. A convex relaxation barrier to tight robustness verification of neural networks//Proceedings of the Advances in Neural Information Processing Systems, 2019: 9832-9842.

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

[69] Mirman M, Singh G, Vechev M. A provable defense for deep residual networks. arXiv preprint arXiv:1903.12519, 2019.

[70] Croce F, Hein M. Provable robustness against all adversarial lp-perturbation for p>=1//Proceedings of the International Conference on Learning Representations, 2020.

[71] Croce F, Andriushchenko M, Hein M. Provable Robustness of ReLU networks via Maximization of Linear Regions//Proceedings of the 22nd

International Conference on Artificial Intelligence and Statistics, 2019: 2057-2066.

[72] Ruan W, Huang X, Kwiatkowska M. Reachability analysis of deep neural networks with provable guarantees. arXiv preprint arXiv:1805.02242, 2018.

[73] Weng T-W, Zhang H, Chen P-Y, et al. On extensions of clever: A neural network robustness evaluation algorithm//Proceedings of the 2018 IEEE Global Conference on Signal and Information Processing, 2018: 1159-1163.

[74] Latorre F, Rolland P, Cevher V. Lipschitz constant estimation for neural networks via sparse polynomial optimization//Proceedings of the International Conference on Learning Representations, 2020.

[75] Lasserre J B. An introduction to polynomial and semi-algebraic optimization. 52. Cambridge University Press, 2015.

[76] Weisser T, Lasserre J B, Toh K-C. Sparse-BSOS: a bounded degree SOS hierarchy for large scale polynomial optimization with sparsity. Mathematical Programming Computation, 2018, 10(1): 1-32.

[77] Li B, Chen C, Wang W, et al. Certified Adversarial Robustness with Additive Noise//Proceedings of the Advances in Neural Information Processing Systems, 2019: 9459-9469.

[78] Tsuzuku Y, Sato I, Sugiyama M. Lipschitz-margin training: Scalable certification of perturbation invariance for deep neural networks//Proceedings of the Advances in Neural Information Processing Systems, 2018: 6541-6550.

[79] Pinot R, Meunier L, Araujo A, et al. Theoretical evidence for adversarial robustness through randomization//Proceedings of the Advances in Neural Information Processing Systems, 2019: 11838-11848.

[80] Lee G-H, Yuan Y, Chang S, et al. Tight certificates of adversarial robustness for randomly smoothed classifiers//Proceedings of the Advances in Neural Information Processing Systems, 2019: 4911-4922.

[81] Xie C, Wang J, Zhang Z, et al. Mitigating Adversarial Effects Through Randomization//Proceedings of the International Conference on Learning Representations, 2018.

[82] Dhillon G S, Azizzadenesheli K, Lipton Z C, et al. Stochastic activation pruning for robust adversarial defense. arXiv preprint arXiv:1803.01442, 2018.

[83] Salman H, Li J, Razenshteyn I, et al. Provably robust deep learning via adversarially trained smoothed classifiers//Proceedings of the Advances in Neural Information Processing Systems, 2019: 11289-11300.

[84] Dvijotham K, Hayes J, Balle B, et al. A framework for robustness certification of smoothed classifiers using f-divergences//Proceedings of the International Conference on Learning Representations, 2020.

[85] Salman H, Sun M, Yang G, et al. Black-box Smoothing: A Provable Defense for Pretrained Classifiers. arXiv preprint arXiv:2003.01908, 2020.

[86] Jia J, Cao X, Wang B, et al. Certified Robustness for Top-k Predictions against Adversarial Perturbations via Randomized Smoothing. arXiv preprint arXiv:1912.09899, 2019.

[87] Wang B, Cao X, Gong N Z. On Certifying Robustness against Backdoor Attacks via Randomized Smoothing. arXiv preprint arXiv:2002.11750, 2020.

[88] Weber M, Xu X, Karlas B, et al. RAB: Provable Robustness Against Backdoor Attacks. arXiv preprint arXiv:2003.08904, 2020.

[89] Mohapatra J, Ko C-Y, Liu S, et al. Rethinking Randomized Smoothing for Adversarial Robustness. arXiv preprint arXiv:2003.01249, 2020.

[90] Sunaga T. Theory of an interval algebra and its application to numerical analysis. RAAG memoirs, 1958, 2(29-46): 209.

[91] Chiang P-Y, Ni R, Abdelkader A, et al. Certified defenses for adversarial patches. arXiv preprint arXiv:2003.06693, 2020.

[92] Zhang H, Chen H, Xiao C, et al. Towards stable and efficient training of verifiably robust neural networks. arXiv preprint arXiv:1906.06316, 2019.

[93] Wang S, Pei K, Whitehouse J, et al. Formal security analysis of neural networks using symbolic intervals//Proceedings of the 27th USENIX Security Symposium, 2018: 1599-1614.

[94] Wang S, Pei K, Whitehouse J, et al. Efficient formal safety analysis of neural networks//Proceedings of the Advances in Neural Information Processing Systems, 2018: 6367-6377.

[95] Mangal R, Nori A V, Orso A. Robustness of neural networks: a probabilistic and practical approach//Proceedings of the 2019 IEEE/ACM 41st International Conference on Software Engineering: New Ideas and Emerging Results, 2019: 93-96.

[96] Weng L, Chen P-Y, Nguyen L, et al. PROVEN: Verifying Robustness of Neural Networks with a Probabilistic Approach//Proceedings of the International Conference on Machine Learning, 2019: 6727-6736.

[97] Fazlyab M, Morari M, Pappas G J. Probabilistic Verification and Reachability Analysis of Neural Networks via Semidefinite Programming. arXiv preprint arXiv:1910.04249, 2019.

[98] Webb S, Rainforth T, Teh Y W, et al. A statistical approach to assessing neural network robustness. arXiv preprint arXiv:1811.07209, 2018.

[99] Wang Y-S, Weng T-W, Daniel L. Verification of Neural Network Control Policy Under Persistent Adversarial Perturbation. arXiv preprint arXiv:1908.06353, 2019.

[100] Carr S, Jansen N, Topcu U. Verifiable RNN-Based Policies for POMDPs Under Temporal Logic Constraints. arXiv preprint arXiv:2002.05615, 2020.

[101] Wang B, Shi Z, Osher S. ResNets Ensemble via the Feynman-Kac Formalism to Improve Natural and Robust Accuracies//Proceedings of the Advances in Neural Information Processing Systems, 2019: 1655-1665.

[102] Balunovic M, Baader M, Singh G, et al. Certifying Geometric Robustness of Neural Networks//Proceedings of the Advances in Neural Information Processing Systems, 2019: 15287-15297.

[103] Jia J, Wang B, Cao X, et al. Certified Robustness of Community

Detection against Adversarial Structural Perturbation via Randomized Smoothing. arXiv preprint arXiv:2002.03421, 2020.

[104] Guidotti D, Leofante F, Pulina L, et al. Verification of Neural Networks: Enhancing Scalability through Pruning. arXiv preprint arXiv:2003.07636, 2020.

[105] Steinhardt J, Koh P W W, Liang P S. Certified defenses for data poisoning attacks//Proceedings of the Advances in Neural Information Processing Systems, 2017: 3517-3529.

[106] Liu X, Cheng M, Zhang H, et al. Towards robust neural networks via random self-ensemble//Proceedings of the European Conference on Computer Vision, 2018: 369-385.

[107] Sehwag V, Wang S, Mittal P, et al. On pruning adversarially robust neural networks. arXiv preprint arXiv:2002.10509, 2020.

[108] Xiao K Y, Tjeng V, Shafiullah N M, et al. Training for faster adversarial robustness verification via inducing relu stability. arXiv preprint arXiv:1809.03008, 2018.

[109] Du T, Ji S, Li J, et al. Sirenattack: Generating adversarial audio for end-to-end acoustic systems//Proceedings of the 15th ACM Asia Conference on Computer and Communications Security. 2020: 357-369.

[110] Li X, Ji S, Han M, et al. Adversarial examples versus cloud-based detectors: A black-box empirical study. IEEE Transactions on Dependable and Secure Computing, 2019.

[111] Ling X, Ji S, Zou J, et al. Deepsec: A uniform platform for security analysis of deep learning model//Proceedings of the 2019 IEEE Symposium on Security and Privacy. IEEE, 2019: 673-690.

[112] Li J, Ji S, Du T, et al. TextBugger: Generating Adversarial Text Against Real-world Applications//Proceedings of the 26th Annual Network and Distributed System Security Symposium. 2019.

[113] Li J, Du T, Ji S, et al. Textshield: Robust text classification based on multimodal embedding and neural machine translation//Proceedings of the 29th USENIX Security Symposium. 2020: 1381-1398.

[114] Shi C, Ji S, Liu Q, et al. Text captcha is dead? a large scale deployment and empirical study//Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security. 2020: 1391-1406.

[115] Zhao B, Weng H, Ji S, et al. Towards evaluating the security of real-world deployed image captchas//Proceedings of the 11th ACM Workshop on Artificial Intelligence and Security. 2018: 85-96.

[116] Ji S L, Du T Y, Li J F, Shen C, Li B. Security and privacy of machine learning models: A survey. Journal of Software, 2021, 32(1):41-67 (in Chinese)

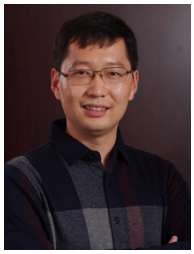
纪守领, 杜天宇, 李进锋, 沈超, 李博. 机器学习模型安全与隐私研究综述. 软件学报, 2021, 32(1):41-67.

[117] Ji S L, Li J F, Du T Y, Li B. Survey on Techniques, Applications and Security of Machine Learning Interpretability. Journal of Computer Research and Development, 2019, 56(10): 2071-2096 (in Chinese)

纪守领, 李进锋, 杜天宇, 李博. 机器学习模型可解释性方法、应用与安全研究综述. 计算机研究与发展, 2019, 56(10): 2071-2096.

[118] Li X R, Ji S L, Wu C M, Liu Z G, Deng S G, Cheng P, Yang M, Kong XW. Survey on deepfakes and detection techniques. Journal of Software, 2021, 32(2):496−518 (in Chinese)

李旭嵘, 纪守领, 吴春明, 刘振广, 邓水光, 程鹏, 杨珉, 孔祥维. 深度伪造与检测技术综述. 软件学报, 2021, 32(2):496−518.



JI Shou-Ling, born in 1986, ZJU 100-Young Professor in the College of Computer Science and Technology at Zhejiang University and Research Faculty in the School of Electrical and Computer Engineering at Georgia Institute of Technology. He received a Ph.D. in Electrical and Computer Engineering from Georgia Institute of Technology and a Ph.D. in Computer Science from Georgia State University. His current research interests include Big Data Security and Privacy, Big Data Driven Security and Privacy, and Adversarial Learning. He also has interests in Graph Theory and Algorithms and Wireless Networks.

DU Tian-Yu, born in 1996, she is a Ph.D. student in the College of Computer Science and Technology at Zhejiang University, under the supervision of Prof. Shouling Ji. She received her BS degree from Xiamen University in 2017. Her research interests include big data driven security, adversarial learning, and AI security.

DENG Shui-Guang, he is a full professor at the College of Computer Science and Technology in Zhejiang University. He received the BS and PhD both in Computer Science from Zhejiang University in 2002 and 2007, respectively. His research interests include Service Computing, Edge Computing, Business Process Management and Big Data.

CHENG Peng, he received the B.E. degree in Automation, and the Ph.D. degree in Control Science and Engineering in 2004 and 2009 respectively, both from Zhejiang University. From 2012 to 2013, he worked as Research Fellow in Information System Technology and Design Pillar, Singapore University of Technology and Design. His research interests include networked sensing and control, cyber-physical systems, control system security.

SHI Jie, he is a principle researcher in Huawei Singapore Research Centre. His research interests include data security, blockchain, AI security and privacy.

YANG Min, he is a professor in the school of Computer Science at Fudan University. His research interests include network security, malicious code detection, vulnerability analysis, AI security, blockchain security, and system security.

LI Bo, born in 1989, she is an assistant professor in the department of Computer Science at University of Illinois at Urbana–Champaign, and is a recipient of the Symantec Research Labs Fellowship. Her research focuses on both theoretical and practical aspects of security, machine learning, privacy, game theory, and blockchain. She is a member of IEEE and ACM.

Background

This research belongs to AI Security area, focusing on the research progress of model robustness certification in the field of adversarial machine learning. Current deep neural networks are known to be vulnerable to malicious manipulations, such as adversarial examples that force target deep neural networks to misbehave. In recent years, a plethora of work has focused on constructing adversarial examples in various domains. The phenomenon of adversarial examples demonstrates the inherent lack of robustness of deep neural networks, which limits their use in security-critical applications.

In order to build a safe and reliable deep learning system and eliminate the potential security risks of deep learning models in real-world applications, the security issue of deep learning has attracted extensive attention from academia and industry. Thus far, intensive research has been devoted to improving the robustness of DNNs against adversarial attacks. Unfortunately, most defenses are based on heuristics and thus lack any theoretical guarantee, which can often be defeated or circumvented by more powerful attacks. Therefore, defenses only showing empirical success against attacks, are difficult to be concluded robust. Unfortunately, most defenses are based on heuristics and thus lack any theoretical guarantee, thus the concept of certifiable robustness is proposed to provide guaranteed robustness by formally verifying whether a given region surrounding a data point admits any adversarial example. A large number of researchers have conducted in-depth research on the model robustness certification from the perspective of complete and incomplete, and proposed a series of certification methods. These methods can be generally categorized as exact certification methods and relaxed certification methods.

In this survey, we review current challenges of model robustness certification problem, systematically and scientifically summarize existing research work, and clarify the advantages and disadvantages of current research. Finally, we explore future research directions of model robustness certification research.

This work was partly supported by the Zhejiang Provincial Natural Science Foundation for Distinguished Young Scholars under No. LR19F020003, the National Key Research and Development Program of China under No. 2020YFB2103802, NSFC under No. 61772466, U1936215, and U1836202, and the Fundamental Research Funds for the Central Universities (Zhejiang University NGICS Platform).