北京大学

博士研究生学位论文

题目：基于验证技术的深度神经网络可信保障

姓　　名：	薛骁勇
学　　号：	1901110055
院　　系：	数学科学学院
专　　业：	应用数学
研究方向：	程序理论、软件形式化方法
导师姓名：	孙猛 教授

☑ 学术学位　　☐ 专业学位

二〇二四 年 五 月

版权声明

摘要

随着计算资源的丰富和算法的创新，深度神经网络在过去十年中取得了显著的进展和突破。在计算机视觉、自然语言处理、语音识别、推荐系统等领域，深度神经网络有着卓越的表现，其准确性和效率可以媲美甚至超越人类。由于其优异的表现，深度神经网络也被应用到自动驾驶、无人机控制、医疗诊断等许多安全攸关系统中做关键决策。与此同时，以对抗样本为代表的一系列隐藏风险被发现，这揭示了深度神经网络在安全性、鲁棒性上的隐患，引起了公众对于深度神经网络是否可信的担忧。如何保障深度神经网络的可信性已经备受关注，成为了其在安全攸关领域应用的关键挑战。

形式化验证技术在传统软件系统的可信性保障方面得到了广泛的认可和使用。其通过精确的定义和严格的证明来验证软件是否符合预定规范，能够发现常规测试无法发现的错误。由于深度神经网络和传统软件在结构和行为上有着显著的差别，使用验证技术对深度神经网络的可信性进行保障面对许多挑战：深度神经网络是数据驱动的生成模式，缺乏明确的决策逻辑，难以确定单个或一组神经元的语义；深度神经网络的算子种类繁多，验证方法难以通用，需要对每个算子专门进行设计；深度神经网络规模庞大，对验证方法的效率和可扩展性提出了很高的要求。这些特点使得传统软件系统的形式化验证技术无法直接应用于深度神经网络。本文基于深度神经网络自身的特点，考虑其所面临的可信性问题，对深度神经网络的形式化验证方法进行研究。对含有不同算子的深度神经网络，本文提出了相应的形式化验证方法；对主流的分枝定界验证框架，本文设计了高效分枝策略以提升验证的效率。

本文的第一部分对使用 ReLU 激活函数的深度神经网络提出了一种基于多神经元松弛的验证框架，它可以用于验证输入扰动在一般范数空间的性质。现有的松弛方法在单个神经元上进行，验证精度的提升遇到了瓶颈。此外，现有的验证方法在深度神经网络的输入层会引入可行区域外的输入，这会导致显著的误差。本文中的验证框架提出了多神经元松弛方法，对同层的神经元划分为不同组，在组内使用多面体抽象获得多个神经元之间的线性关系，并且设计了松弛选择方法，从而选择更严格的凸松弛以提高验证精度。为了排除可行区域外的输入，本文中还提出了一种区域裁剪方法，通过准确地求解一个约束优化问题进一步提高验证精度。

本文的第二部分提出了适用于含 sigmoid-like 激活函数神经网络的分枝定界验证框架。由于 sigmoid-like 激活函数的非线性特性，此类神经网络的验证依赖于线性近似方法，这不可避免地会引入误差。分枝定界技术可以不断精化线性近似的结果，达到更高的验证精度。在此框架中设计了针对 sigmoid-like 激活函数的神经元分裂方法

和分枝策略。神经元分裂方法基于父问题的线性松弛和激活函数的凹凸性，将非线性激活函数分割为多个片段，为每个片段计算线性上界和线性下界，保证了分枝定界验证过程的单调性。本文中还提出了适配该验证框架的分枝策略。该策略可以有效减少分枝定界搜索树的大小，提高验证效率。实验表明，该框架的验证结果比现有适用于 sigmoid-like 神经网络验证方法的结果更加精确。

本文的第三部分提出了一种适用于多种神经网络的分枝策略。分枝策略是分枝定界验证框架中的一项关键组成部分，它决定了问题的可行区域如何分割。好的分枝策略可以减少验证过程中需要探索的分枝数量，从而提高验证效率。本文中的分枝策略根据分裂神经元所产生子问题相对父问题的提升对每个神经元打分。分裂分数高的神经元所产生的子问题更加可能被验证，此神经元为更好的分枝决策。为了保证分枝策略的效率，本文将定界验证算法所求得的父问题最小值拓展到神经网络的各层，并以此来估计子问题相对父问题的提升。本文中的分枝策略还包括界外补偿和分数截断技术，这些技术对每个神经元的分数进行了修正。此外，我们证明了某些分枝选择产生的一些子问题可以利用父问题的最优解直接得到最小值，这减少了对验证算法的调用次数。实验结果表明，该分枝策略可以有效减少验证过程中产生的分枝数量和验证时间，提升验证效率。

本文的研究内容为面向深度神经网络可信性的形式化验证技术，给出了一套基于线性松弛和分枝定界技术的深度神经网络可信性保障框架，并对具备不同算子的深度神经网络进行了专门优化以提升验证的精度与效率。该框架为深度神经网络的可信性问题提供了严谨且高效的解决方案，对深度神经网络在安全攸关领域的应用及可信深度神经网络的发展有重要意义。

关键词：深度神经网络，形式化验证，可信保障，分枝定界法，分枝策略

Verification Based Trustworthiness Assurance for Deep Neural Networks

Xiaoyong Xue (Applied Mathematics)
Directed by Prof. Meng Sun

ABSTRACT

With the abundance of computing resources and the innovation of neural network algorithms, deep neural networks have made significant progress and breakthroughs in the past decade. In many fields, such as computer vision, natural language processing, speech recognition, recommendation systems, deep neural networks have demonstrated outstanding performance, with accuracy and efficiency that can match or even outperform human beings. Due to their excellent performance, deep neural networks have also been applied in many safety-critical systems to make crucial decisions, such as autonomous driving, unmanned aerial vehicle control, medical diagnosis, etc. Meanwhile, a series of risks represented by adversarial examples have been discovered. This reveals the potential risks in the security and robustness of deep neural networks, raising public concerns about their trustworthiness. The trustworthiness issues of deep neural networks have become a significant barrier to their application in safety-critical fields.

Formal verification techniques have been widely recognized and used to ensure the trustworthiness of traditional software systems. These techniques verify whether software conforms to predefined specifications through precise definitions and rigorous proofs, and reveal errors that conventional testing techniques fail to detect. However, due to the significant differences in structure and behavior between deep neural networks and traditional software, there exists many challenges in using verification techniques to ensure the trustworthiness of deep neural networks. Deep neural networks are data-driven models generated by training algorithms. Due to the lack of explicit decision logic, it is difficult to determine the semantics of individual or groups of neurons. There exists different types of operators in deep neural networks, making it hard for verification methods to be universal, and thus requiring specialized designs for each operator. The massive scale of deep neural networks requires highly efficient and scalable verification methods. These characteristics make that formal verification techniques for traditional

software systems unable to be directly applied to deep neural networks. Therefore, based on the characteristics of deep neural networks and the robustness and trustworthiness issues, this thesis studies formal verification methods for deep neural networks. This thesis proposes dedicated formal verification methods for deep neural networks with different operators. For the mainstream branch and bound verification framework, this thesis designs efficient branching strategies to improve the verification efficiency.

The first part of this thesis proposes a verification framework for deep neural networks with ReLU activation functions, which is based on multi-neuron relaxation and can be used to verify properties with input perturbations in general norm spaces. The current relaxation methods for individual neuron have reached a bottleneck in improving verification accuracy. Furthermore, existing verification methods involve inputs that lie outside the feasible region at the input layer, resulting in notable errors. The verification framework in this part presents a multi-neuron relaxation method that divides neurons within the same layer into several groups and uses polyhedral abstraction to establish linear relationships between multiple neurons within a group. In the presence of multiple candidate relaxations, we propose a relaxation selection method to select tighter convex relaxations. In order to exclude inputs outside the feasible region, we also propose a region clipping method, which solves a constrained optimization problem to improve the verification ability.

The second part of this thesis presents a branch-and-bound verification framework, which is suitable for neural networks with sigmoid-like activation functions. Due to the nonlinear nature of sigmoid-like activation functions, verification of such neural networks mostly relies on linear approximation methods, which inevitably introduce errors and lead to imprecise results. The branch-and-bound technique iteratively refines the results of linear approximations and is able to achieve higher precision. This framework contains a neuron splitting method and a branching strategy. The neuron splitting method divides the nonlinear activation function into several segments based on the linear relaxation of the parent problem and the concavity of the activation function, and computes a linear upper bound and a linear lower bound for each segment, which ensures the monotonicity of the branch and bound verification process. Additionally, we propose a dedicated branching strategy for this verification framework. This strategy can effectively reduce the size of the branch-and-bound search tree, thereby improving verification efficiency. Experiments show that the verification results obtained with our verification framework are more precise compared to those from existing state-of-the-art verification methods for sigmoid-like neural networks.

The third part of this thesis proposes a branching strategy that can be applied to various neural networks. Branching strategy is a critical component in the branch and bound verification framework, determining how the feasible region of the problem is divided. A good branching strategy can reduce the number of branches that need to be explored during the verification process, thereby improving verification efficiency. The branching strategy assign a score to each neuron based on the improvement of the sub-problems relative to the parent problem. Neurons with higher scores are more likely to produce sub-problems that can be directly verified, making them better candidates for branching decisions. To ensure the efficiency of the branching strategy, we extend the optimal solution obtained by the bounding verification algorithm for the parent problem to all layers in the neural network, which is used to estimate the improvement of the sub-problems relative to the parent problem. Additionally, the branching strategy includes out-of-bound compensation and score truncation techniques, which adjust the score for each neuron. We also prove that some sub-problems generated by certain branch choices can be directly solved, thereby reducing the number of calls to the bounding algorithm. Experimental results show that this branching strategy effectively improve verification efficiency by reducing the verification time and the number of branches produced during the verification process.

This thesis makes contributions to formal verification techniques for trustworthiness assurance of deep neural networks, presenting a deep neural network verification framework based on linear relaxation and branch and bound techniques. It is specially optimized for deep neural networks with different operators to improve the precision and efficiency of verification. This framework provides rigorous and efficient solutions to the safety, security and trustworthiness problems of deep neural networks, and is important to the application of deep neural networks in safety-critical areas and the development of trustworthy deep neural networks.

KEY WORDS: Deep Neural Network, Formal Verification, Trustworthiness Assurance, Branch and Bound, Branching Strategy

Contents

攻读博士期间发表或接收的论文及其他成果

个人简介

薛骁勇，男，2015 年 9 月至 2019 年 6 月就读于北京大学数学科学学院，2019 年 6 月获得信息与计算科学学士学位。2019 年 9 月至 2024 年 6 月于北京大学数学科学学院攻读博士学位，研究方向为程序理论、软件形式化方法。

已发表论文

1. **Xiaoyong Xue**, Xiyue Zhang, Meng Sun. kProp: Multi-neuron Relaxation Method for Neural Network Robustness Verification. In: *Proceedings of FSEN 2023*, LNCS 14155, Tehran, Iran: Springer 2023, pp. 142-156.
2. **Xiaoyong Xue** and Meng Sun. Branch and Bound for Sigmoid-Like Neural Network Verification. In: *Proceedings of ICFEM 2023*, LNCS 14308, Brisbane, QLD, Australia: Springer 2023, pp. 137-156.
3. **Xiaoyong Xue** and Meng Sun. Optimal Solution Guided Branching Strategy for Neural Network Branch and Bound Verification. Accepted by ICECCS2024, Limassol, Cyprus.
4. Weidi Sun, **Xiaoyong Xue**, Yuteng Lu, and Meng Sun. HashC: Making DNNs' Coverage Testing Finer and Faster. In: *Proceedings of SETTA 2022*, LNCS 13649, Virtual, Online: Springer, 2022, pp. 3-21.
5. Weidi Sun, **Xiaoyong Xue**, Yuteng Lu, Jia Zhao and Meng Sun. HashC: Making deep learning coverage testing finer and faster. In: *Journal of Systems Architecture*, 144(2023).
6. Xiangyu Li, Yihao Zhang, Xiaokun Luan, **Xiaoyong Xue** and Meng Sun. MedTiny: Enhanced Mediator Modeling Language for Scalable Parallel Algorithms. In: *Proceedings of QRS-C 2023*, Chiang Mai, Thailand: IEEE, 2023, pp. 451-460.
7. **薛骁勇**, 孙猛. Mediator 的概率扩展 [J]. 计算机工程与科学, 2020, 42(08):1367-1373.

已投稿论文

1. **Xiaoyong Xue**, Xiyue Zhang, Meng Sun. kProp: Multi-Neuron Relaxation Based Verification for Neural Network with General Activation Functions. Manuscripts.
2. Xiyue Zhang, **Xiaoyong Xue**, Xiaoning Du, Xiaofei Xie, Meng Sun and Yang Liu. Runtime Backdoor Detection for Federated Learning via Representational Dissimilarity Analysis. Manuscripts.

获奖情况

1. 2023 年 6 月. 数学学院学院奖学金
2. 2023 年 9 月. 优秀科研奖
3. 2023 年 9 月. 五四奖学金

致谢

桌面上的台历翻开了新的一页，四季的时钟又转动到一个夏天。在博士生涯的最后一小段时光里，我希望借此机会对我过去五年甚至更长的旅途做一个简短的回顾，并对在我前行路上给予我指导、帮助和鼓励的老师、同学以及亲友献上我最真挚的谢意。

首先，我衷心地感谢我的导师孙猛教授。与孙老师的初识是本科阶段的讨论班，那时的我初来乍到，对形式化方法领域一无所知，孙老师给予了我足够的耐心、循序渐进的指导，使我慢慢地找到了学术的方向，让我从科研新手逐渐蜕变成了一个更成熟的科研工作者。在生活中，每当我倍感压力或者迷茫无助时，孙老师都会鼓励引导我，让我有了面对困难与挫折的勇气。孙老师待人宽容温和，治学严谨，关心学生的需要，重视学生的发展，是我一直以来的目标和榜样。虽然博士生活即将告一段落，但师生情谊将久久长存。

我还要感谢教授我专业知识以及给予我指导和交流的老师。感谢在北大教授我博士生专业课程知识的夏壁灿老师、胡振江老师、熊英飞老师、杨建生老师、林作铨老师、牟克典老师、金芝老师、张行功老师、万小军老师。你们的课程是我前行的基础，为我打开了各个领域的大门。感谢曾在讨论班和各种会议上与我有过交流并对我进行指导的张民老师、宋富老师、詹乃军老师、张立军老师、蔡少伟老师、白光冬老师、黄小炜老师、陈立前老师、谢肖飞老师。你们的交流与指导如醍醐灌顶，给了我深刻的启发。

此外我还要感谢在研究工作中与我合作、共同前行的同学朋友。感谢张喜悦师姐在我科研初期对我的帮助。在合作的过程中，她对待科学研究的好奇心和严谨的工作习惯深深影响着我，让我在论文写作方面受益匪浅。感谢孙纬地师兄在深度神经网络测试方面和我进行的讨论，开阔了我的视野。感谢栾晓坤师弟神经网络验证工作中提供的建议和交流，每次交流都能相互启发。

我还要感谢这一路走来在学习上生活中给予我帮助鼓励的同学、朋友。感谢一起交流过的李屹博士、刘艾博士、卢煜腾博士、李昊坤博士、卜昊同学、李翔宇同学、徐紫云同学、吴雨伦同学、张琦同学、冯逸群同学、杨晓宇同学、李忆同学。感谢我的室友侯浩杰同学在生活中的帮助。感谢我的朋友辛天屹、刘禹、刘逸涵，感谢你们平时对我点点滴滴的帮助，陪我度过了这段难忘的时光。

最后我衷心地感谢我的父母、家人，感谢他们对我的理解和一如既往的支持，在我迷茫时给我鼓励，在我受挫时给我力量。感谢赵天琪女士对我科研、学习和生活上

的鼓励和帮助, 陪伴我度过博士生涯。