博士研究生学位论文

题目：深度学习系统的覆盖测试与全局验证

姓　　名：孙纬地

学　　号：1801110050

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：孙猛 教授

二〇二三 年 六 月

版权声明

摘要

近年来，深度学习系统开始在越来越多的领域得到应用。此类系统利用神经网络进行搜索、推荐、决策、特征提取，并在计算机视觉、自然语言处理等诸多前沿领域取得了可以媲美甚至超越人类的性能。然而，深度学习系统在医疗诊断、自动驾驶等安全攸关领域的部署也引发了公众对其正确性、鲁棒性等方面的忧虑。尤其是随着对抗攻击被发现以及一系列令人痛心的事故的发生，业界开始认识到深度学习系统与传统软硬件系统一样，面临严峻的安全可信问题，亟待更严格、系统的方法来保障其在安全攸关领域的应用。

深度学习系统与传统软硬件系统行为具有显著区别，为其可信性保障带来了巨大的挑战。例如，深度学习系统本质上遵循数据驱动的编程范式，缺乏显式的控制流；深度学习系统具有涌现性，系统整体行为的复杂性远远大于其组成单元行为的简单累加；深度学习系统往往具有庞大的输入空间，而对抗反例普遍分布在输入空间中，难以被传统的验证方法全部找出，等等。以上这些特点使得传统软硬件系统的认证方法往往难以直接迁移到深度学习系统上，严重限制了现有深度学习可信性保障方法的可用性。本论文从深度学习系统自身的特性出发，对深度学习系统的测试和形式化验证方法进行研究。其中测试是对系统进行可信性分析最普遍的轻量级手段，在系统部署应用前能够有效暴露其中潜在的问题和漏洞，以保障系统的安全性；而形式化验证是与测试互补的一种重量级技术，能为系统提供完备的可信性保障。论文包括如下三部分内容：深度学习系统覆盖测试的有效性、神经网络覆盖准则的改进、及前馈神经网络的全局鲁棒性验证。

论文第一部分探索了深度学习系统覆盖测试标准在度量测试充分性和提升深度学习系统鲁棒性方面的有效性，并提出了新的覆盖测试标准。受传统软件工程测试方法的启发，人们提出了各种基于不同覆盖标准的测试方法来保证深度学习系统的安全性。但在相关研究中，覆盖测试标准在对抗攻击、寻找神经网络漏洞等方面的有效性受到质疑。本文指出了两个适合覆盖测试标准的应用领域：1）评估不同测试集的测试充分性，2) 引导数据增强来提升深度学习系统的鲁棒性，并从这两个方面的实验表现评价了覆盖测试标准的性能。评估结果表明覆盖测试标准在这两个方面的有效性，且我们提出的新神经元覆盖标准在这两方面都优于其他主流的覆盖测试标准。

论文的第二部分提出了加速并细化深度学习系统覆盖测试的框架。已有的深度学习系统覆盖测试标准要么不够精确，无法捕捉神经网络的微妙行为，要么时间代价过高，无法部署在大规模的神经网络上，很难平衡测试充分性评价的质量和效率。此外，

主流的覆盖测试标准在测试套件规模上缺乏“可扩展性”，当评估的测试套件规模过大或过小时，其性能都不符合深度学习系统的测试实践。在本文中我们提出了使用哈希加速的组合覆盖深度学习系统测试框架。该框架利用哈希加密函数对激活状态分析进行加速，同时赋予主流的覆盖标准传递性和组合性以细化其评估粒度并提升其可扩展性。框架将组合覆盖测试的时间复杂度从多项式时间降低到线性时间，从而能够部署在更大规模神经网络上，并且能获得更敏感的测试充分性评估能力。

论文的第三部分提出了可变粒度的深度学习系统覆盖测试标准。覆盖测试标准被应用于评估深度学习系统测试充分性、寻找极端情况、指导测试样例的选择、辅助数据增强等领域。不同的用途对覆盖测试标准的粒度提出了不同的要求，例如评估测试充分性需要粒度尽可能细的覆盖测试标准，而指导测试样例选择需要覆盖测试标准以较粗的粒度给出少量的高价值候选样例。为了给不同种类任务提供通用的覆盖测试标准，本文提出了可变粒度深度学习系统覆盖测试标准 HeatC。该标准从神经网络中提取基于类激活图的特征，并聚类特征来生成测试目标。实验表明，HeatC 在评估测试套件的充分性和从无标注数据集中挑选高价值测试样例方面的表现均优于现有主流覆盖标准。

论文的第四部分提出了全局鲁棒性可验证的深度学习框架。现有的深度学习对抗攻击技术是不完备的，难以在无法找到对抗反例时保障神经网络本身的安全性。而现有的深度学习验证工作多集中于局部鲁棒性，例如指定输入空间中的可行域，分析其在输出空间中的可达性。对于主流的深度学习任务来说，在高维输入空间中指定可行域本身就是“预言家难题”，因为如果我们能在输入空间中以某种简单约束的形式划出某个类别的可行域，那我们就不需要这个深度学习系统。针对这一问题，我们开发了一个深度学习系统全局鲁棒性验证框架 DeepGlobal，该框架包含一个通过符号执行寻找网络潜在边界的规则生成器，以及一个能将规则生成代价降低到多项式时间的神经网络架构。DeepGlobal 从生成的潜在边界中选择神经网络在执行时真实生效的边界，并进一步在边界上寻找非噪声的输入，从而发现会被对抗攻击威胁的边界。

本论文的研究内容为深度学习系统的安全分析及针对对抗攻击的鲁棒性验证技术。给出了一套基于神经网络覆盖测试和全局鲁棒性验证技术的深度学习系统可信保障框架，通过测试与形式化验证的结合，为深度学习系统面临的可信性、安全性问题提供具有高度可扩展性的解决方案，对深度学习系统在安全攸关领域的应用及可信深度学习的发展有重要意义。

关键词：深度学习系统，覆盖测试，全局验证

Coverage Testing and Global Verification of Deep Learning Systems

Weidi Sun (Applied mathematics)
Directed by: Prof. Meng Sun

ABSTRACT

In recent years, deep learning (DL) systems have been applied in more and more fields. Such systems use neural network for search, recommendation, decision, feature extraction, etc., and have achieved performance comparable to or even surpassing human beings in many frontier fields such as computer vision and natural language processing. However, the deployment of DL systems in safety-critical areas such as medical diagnosis and autonomous driving has raised public concerns about its correctness and robustness. Especially with the discovery of adversarial attacks and a series of distressing accidents, the industry has realized that DL systems, like traditional hardware and software systems, face serious security and trustworthiness problems. Rigorous and systematic methods are urgently needed to ensure DL systems' application in security-critical fields.

The behaviors of DL systems are significantly different from those of traditional software and hardware systems, which brings great challenges to DL systems' reliability. For example, DL systems inherently follow a data-driven programming paradigm and lack explicit control flow; DL systems are emergent, and the complexity of their overall behavior is much greater than the simple accumulation of their unit-level behaviors' complexity; DL systems have huge input space containing pervasively distributed adversarial examples, and traditional verification methods can hardly find all these adversarial examples etc. The aforementioned challenges make it difficult to directly migrate the certification methods of traditional software and hardware systems to DL systems, and seriously limit the availability of the existing DL trustworthiness assurance methods. Therefore, this thesis aims to investigate testing and formal verification of the DL systems based on their own characteristics. Testing is the most common light-weight method for trustworthiness guarantee of large-scale DL systems. It can effectively expose the potential problems and vulnerabilities before the deployment, so as to guarantee the trustworthiness of systems. Formal verification is a heavy-weight method which

is complementary to testing. It can provide complete trustworthiness assurance for DL systems. This thesis includes the following three parts: the validity of coverage testing for DL systems, the improvement of neural networks' coverage criteria, and the global robustness verification of feedforward neural networks.

The first part of this thesis explores the validity of DL coverage criteria in two aspects, measuring test adequacy and improving the robustness of DL systems, and proposes a new coverage testing criterion. Inspired by the traditional software engineering testing, testing methods based on various coverage criteria have been proposed to ensure the safety of DL systems. However, the validity of coverage criteria in the adversarial attack, finding vulnerabilities, and other applications has been questioned in related researches. This thesis points out two areas suitable for coverage criteria: 1) evaluating test adequacy of different test sets, 2) guiding data augmentation to improve the robustness of DL systems, and evaluates the performance of coverage criteria via the experiment in these two aspects. The evaluation results show the validity of coverage criteria in these two areas, and our novel coverage criterion is superior to other mainstream criteria in these two aspects.

The second part of this thesis proposes a framework for accelerating and refining coverage testing. Existing coverage criteria are either not fine enough to capture the subtle behavior of neural networks, or too time-consuming to be deployed on large-scale neural networks, which can hardly balance the quality and efficiency of test adequacy evaluation. In addition, some mainstream coverage criteria lack "scalability" regarding test suite size. Their performance does not conform with DNN testing practice when the scale of the evaluated test suite is too big or small. In this thesis, a combinatorial coverage testing framework with hash acceleration is proposed. The framework utilizes cryptographic hash functions to speed up the analysis of activation states, and makes mainstream coverage criteria transitive and combinatorial to refine their evaluation granularity and improve their scalability. The framework reduces the time complexity of combinatorial coverage testing from polynomial time to linear time, enabling its deployment on larger-scale neural networks and more sensitive test adequacy evaluation.

The third part of this thesis presents a variable-grained DL coverage criterion. Coverage criteria are applied to many areas, such as evaluating the test adequacy of deep learning systems, finding corner cases, guiding the selection of test samples, assisting data augmentation, etc. Different applications require coverage criteria with different levels of granularity. For example, the coverage criteria for evaluating the test adequacy need to be as fine-grained as possible, while guiding the test sample selection requires the coverage criteria to provide a

small number of high-value candidates at a coarser granularity. To provide a common coverage criterion for different tasks, this thesis proposes a variable-grained DL coverage criterion: HeatC. It extracts class-activation-map-based features from neural networks, and clusters the features to generate test targets. HeatC outperforms existing mainstream coverage criteria in assessing the adequacy of test suites and selecting high-value test samples from unlabeled dataset.

The fourth part of the thesis proposes a DL framework for global robustness verification. The existing DL adversarial attack technologies are incomplete, as they cannot guarantee the safety of neural networks when adversarial examples cannot be found. Meanwhile, existing DL verification works mostly focus on local robustness, such as analyzing the output space reachability of a specified feasible region in input space. For mainstream DL tasks, specifying a feasible region in a high-dimensional input space is an "Oracle Problem", because if we can specify the feasible regions of a certain category in the input space in the form of simple constraints, we do not need the DL systems. To address this problem, a framework for global robustness verification of DL systems named DeepGlobal is developed in this part. DeepGlobal has a rule generator that finds the potential boundaries of the network via symbolic execution, and a neural network architecture that reduces the cost of rule generation to polynomial time. From the generated potential boundaries, DeepGlobal selects the real boundaries taking effect in the execution of the neural network, and searches for non-noise inputs around the real boundaries to find the adversarial dangerous boundaries.

The research contents of this thesis consist of the safety analysis of DL systems and robustness verification techniques against adversarial attacks. A trustworthiness assurance framework based on neural network coverage testing and global robustness verification is proposed, which provides highly scalable solutions to the credibility and safety problems faced by DL systems through the combination of testing and formal verification. It is of great significance to the application of DL systems in safety-critical fields and the development of trustworthy DL systems.

KEY WORDS: Deep Learning Systems,Testing,Coverage Criteria,Formal Verification

Contents

攻读博士期间发表的论文及其他成果

个人简介

孙纬地，男，2014 年 9 月至 2018 年 6 月就读于北京大学元培学院，2018 年 6 月获得信息与计算科学学士学位。2018 年 9 月至 2023 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. **Weidi Sun**, Yuteng Lu, Xiyue Zhang, and Meng Sun. “DeepGlobal: A Framework for Global Robustness Verification of Feedforward Neural Networks”. In: *Journal of Systems Architecture*, 128(2022).
2. **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.
3. **Weidi Sun**, Yuteng Lu, Xiyue Zhang, and Meng Sun. “DeepGlobal: a Global Robustness Verifiable FNN Framework”. In: *Proceedings of SETTA 2021*, LNCS 13071, Virtual, Online: Springer, 2021, pp. 22-39.
4. **Weidi Sun**, Yuteng Lu, and Meng Sun. “Are Coverage Criteria Meaningful Metrics for DNNs?”. In: *Proceedings of IJCNN 2021*, Virtual, Online: IEEE, 2021, pp. 1-8.
5. **Weidi Sun**, and Meng Sun. “PRISM Code Generation for Verification of Mediator Models”. In: *Proceedings of SEKE 2019*, Lisbon, Portugal: KSI Research Inc. and Knowledge Systems Institute Graduate School, 2019, pp. 271-274.
6. Yuteng Lu, **Weidi Sun**, and Meng Sun. “Towards Mutation Testing of Reinforcement Learning Systems”. In: *Journal of Systems Architecture*, 131(2022).
7. Yuteng Lu, Kaicheng Shao, **Weidi Sun**, and Meng Sun. “MTUL: Towards Mutation Testing of Unsupervised Learning Systems”. In: *Proceedings of SETTA 2022*, LNCS 13649, Virtual, Online: Springer, 2022, pp. 22-40.
8. Yuteng Lu, Kaicheng Shao, **Weidi Sun**, and Meng Sun. “RGChaser: A RL-Guided Fuzz and Mutation Testing Framework for Deep Learning Systems”. In: *Proceedings of DSA 2022*, Wulumuqi, China: IEEE, 2022, pp. 12-23.
9. Yuteng Lu, **Weidi Sun**, and Meng Sun. “Mutation Testing of Reinforcement Learning

Systems”. In: *Proceedings of SETTA 2021*, LNCS 13071, Virtual, Online: Springer, 2021, pp. 143-160.

10. Yuteng Lu, **Weidi Sun**, Guangdong Bai, Meng Sun. “DeepAuto: A First Step Towards Formal Verification of Deep Learning Systems”. In *Proceedings of SEKE 2021*, Pittsburgh, PA, United states: KSI Research Inc. and Knowledge Systems Institute, 2021, pp. 172-176.
11. Yi Li, **Weidi Sun**, and Meng Sun. “Mediator: A Component-based Modeling Language for Concurrent and Distributed Systems”. In: *Science of Computer Programming*, 192(2020).

已投稿论文

1. **Weidi Sun**, Xiaoyong Xue, Yuteng Lu, Jia Zhao, and Meng Sun. “HashC: Making DNNs’ Coverage Testing Finer and Faster”. Manuscripts.
2. **Weidi Sun**, Yuteng Lu, Meng Sun. “HeatC: A Variable-grained Coverage Criterion for Deep Learning Systems”. Manuscripts.
3. Yuteng Lu, Kaicheng Shao, **Weidi Sun**, Jia Zhao and Meng Sun. “MTUL: a Mutation Testing Approach of Unsupervised Learning Systems”. Manuscripts.
4. Yuteng Lu, Kaicheng Shao, **Weidi Sun** and Meng Sun. “A RL-Guided Fuzz and Mutation Testing Framework for DL Systems and Its Ecosystem”. Manuscripts.

专利

1. 一种屏幕翻拍检测方法，孙纬地，郭烽，苏晓东，字节跳动（已受理）
2. 一种从像素化明水印图像中还原文本信息的技术，孙纬地，郭烽，苏晓东，字节跳动（已受理）
3. 针对压缩退化明水印图像的复原技术框架，孙纬地，郭烽，苏晓东，字节跳动（已受理）

获奖情况

1. 2022 年 10 月. 斯伦贝谢奖学金
2. 2022 年 9 月. 校级三好学生
3. 2022 年 6 月. 校长奖学金
4. 2020 年 6 月. 学院奖学金

5. 2020 年 6 月. 优秀科研奖
6. 2019 年 6 月. 学院奖学金

参与项目

1. 2022-2025，深度学习系统的可信性保障
2. 2021-2022，大规模深度学习系统形式化建模与验证
3. 2019-2021，高可信深度学习：理论与技术

致谢

在博士生涯的最后十五分之一里，我希望以自己的谢意来总结这五年甚至更加漫长的旅途。

首先，我要由衷地感谢我的导师孙猛教授。在本科的时候加入孙老师的讨论班是一件很幸运的事，这让我从元培学院多到让人眼花缭乱的可选方向中逐渐明确自己的兴趣，也让我师从于一位严谨、宽容、仁厚的学者。在孙老师耐心的指导和帮助下，我开始迈出自己学术生涯的许多第一步：调研第一个领域、遇到第一个难题、发表第一份论文、参加第一次学术会议、进行第一次合作。每一个第一背后都有很多令我难忘的瞬间，那些深夜微信里布满批注的论文，办公室里长谈时窗外的风声，下课后边走边聊的开悟共同成为了我博士生涯中的亮色和感动的一部分。孙老师的言传身教让我明白，学术是自由的，也是严谨的；工作是需要自己辛苦付出的，也是需要与人互通有无的；学术生涯中灵光一闪是凤毛麟角的，而考验耐心和细心的任务是日复一日的；崇高无拘的梦想是第一动力，但勤奋自律的人格才能让人离真理更近。这些道理不仅仅是做博士生的道理，也是生活的道理，为学之师与为人之师，孙老师大概是做了两份工作。

其次，我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查红彬老师、金芝老师、李戈老师、林通老师、林作铨老师、谭营老师、夏壁灿老师、熊英飞老师、杨建生老师、张行功老师。感谢曾经带领我在神经网络和计算机视觉方面入门的马尽文老师。感谢曾在讨论班和各种会议上与我有过交流指导的白光冬老师、黄小炜老师、刘万伟老师、刘杨老师、蒲戈光老师、裘宗燕老师、佘志坤老师、宋富老师、王戟老师、詹乃军老师、张立军老师、Prof. Farhad Arbab、Prof. Grigore Rosu 等各位老师。

此外我还要感谢在研究项目中与我共同合作的同学、朋友。感谢卢煜腾同学在我深度学习安全相关的一系列工作中提供的帮助，卢同学是我科研上最亲密的战友，我们共同讨论、互相勉励走过了近乎五年的时光，愿我们友谊长存。感谢李屹博士在我科研工作的起步阶段为我提供的指导，帮助我对形式化方法及相关工具有了初步的认识。感谢薛骁勇同学为我在神经网络测试工作的实验中提供的帮助以及同我在神经网络安全问题上的探讨。同样感谢栾晓坤同学在我神经网络验证工作中提供的建议以及我们在神经网络安全问题上的交流。

除了在课题上有过合作的同学，我还感谢在学习和生活中帮助过我的同学、朋友。感谢讨论班上与我共同学习交流的张喜悦博士、刘艾博士、Dr. Saqib Nawaz 、张琦同

学、冯逸群同学、卜昊同学、杨晓宇同学、李忆同学。感谢我的室友陈亮同学为我在生活中提供的帮助。感谢我的朋友宋涛、杨中天、池一、王金，是他们让我觉得人不能没有朋友，至少博士生不能。

最后我衷心地感谢我的父母，感谢他们的支持、包容，感谢他们做我溃不成军时的敦刻尔克，感谢他们打来的每一个电话，如同感谢黑夜中的炬火。感谢我的知己郑怡硕女士，她是我的宁静、我的勇气、我的智慧。

在我第一次踏进燕园的时候，未曾想过当时南门树下的丁达尔效应虽随时光红移但仍能照进今早理教的窗子。我特意从我租住的房子赶回这里，在模糊的胶体粒子与仪式中写我的致谢，却知道我不能如这光景般回环。我在燕园的生活占据了到目前为止人生的三分之一，我的梦和思想于此诞生、于此奔涌、或许也于此趋于温和，我计划并终将驶离这里。

“不知我等是狂是愚，唯知一路向前奔驰。”

北京大学学位论文原创性声明和使用授权说明

原创性声明

本人郑重声明：所呈交的学位论文，是本人在导师的指导下，独立进行研究工作所取得的成果。除文中已经注明引用的内容外，本论文不含任何其他个人或集体已经发表或撰写过的作品或成果。对本文的研究做出重要贡献的个人和集体，均已在文中以明确方式标明。本声明的法律结果由本人承担。

论文作者签名：　　　　　日期：　　年　　月　　日

学位论文使用授权说明

（必须装订在提交学校图书馆的印刷本）

本人完全了解北京大学关于收集、保存、使用学位论文的规定，即：

- 按照学校要求提交学位论文的印刷本和电子版本；
- 学校有权保存学位论文的印刷本和电子版，并提供目录检索与阅览服务，在校园网上提供服务；
- 学校可以采用影印、缩印、数字化或其它复制手段保存论文；
- 因某种特殊原因需要延迟发布学位论文电子版，授权学校□一年/□两年/□三年以后，在校园网上全文发布。

（保密论文在解密后遵守此规定）

论文作者签名：　　　　　导师签名：

日期：　　年　　月　　日