



北京大学

博士研究生学位论文

题目：基于认证技术的深度学习系统
可信性保障

姓名：张喜悦

学号：1701110047

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

二〇二二年六月

版权声明

任何收存和保管本论文各种版本的单位和个人，未经本论文作者同意，不得将本论文转借他人，亦不得随意复制、抄录、拍照或以其他方式传播。否则，引起有碍作者著作权之问题，将可能承担法律责任。



摘要

近年来，随着计算资源及人工智能相关技术的飞速发展，深度学习系统在诸多领域得到了广泛应用，此类系统利用神经网络作为决策组件，在处理计算机视觉、机器翻译等复杂任务的准确性和效率上取得了卓越的表现。与此同时，深度学习系统在自动驾驶、医疗诊断等安全攸关领域的部署和应用也引发了公众对该类系统正确性、鲁棒性的担忧。实际上，与传统软硬件系统一样，深度学习系统面临着严峻的安全可信问题，包括决策鲁棒性和决策过程可解释性的缺乏，这些问题成为了它们应用于安全攸关领域的主要障碍。

认证技术在保障传统软件和硬件系统的安全性和可靠性方面发挥了重大作用。然而，相比于传统软硬件系统，深度学习系统具有独特的内部结构和决策逻辑，现有的针对深度学习系统的可信保障方法往往建立在传统软硬件系统认证技术的基础之上，而缺乏对深度学习系统决策逻辑的支持；另一方面，现有的深度学习系统认证技术在可扩展性上的限制使得其难以适应复杂任务场景中的实际需求。因此，本论文旨在基于深度学习系统自身特点，针对该类系统在对抗攻击和后门攻击下的安全性、鲁棒性缺陷，对适用于深度学习系统可信性保障的认证技术进行研究，具体研究包括深度学习系统的抽象模型构建，神经网络的统一逻辑规约，不确定性标准向导的测试用例生成以及分布式学习中客户端的差异分析。

论文的第一部分提出了深度学习系统的抽象模型提取技术。现有的针对深度学习系统的安全分析和验证技术所能处理的系统规模和类型受到很大限制，复杂的计算过程和内部结构设计使得对神经网络的直接分析变得困难。因此研究近似精度高、可扩展性强的抽象模型构建方法可为复杂系统的分析和验证提供基础。本论文中提出了加权自动机提取算法，通过抽象技术构建原系统的状态转移模型，进一步利用所构建的加权自动机对系统的对抗攻击安全进行分析。所提取的加权自动机还可作为原系统的全局解释模型，提高深度学习系统的可解释性。该算法基于目标神经网络的决策置信度语义构建状态空间的离散划分，利用特征向量的语义相似度提取状态转移规则，极大地增强了加权自动机提取算法相对于先进技术在提取精度上的优势。

论文的第二部分提出了神经网络的统一逻辑规范框架。规范框架的设计和开发对于系统在实际场景中的安全部署至关重要，也是对系统进行安全分析的重要前提。形式化逻辑规范不仅可用于描述神经网络的行为和性质，还可以作为展开定理证明和模型检查的基础，保障神经网络行为的安全性。本论文中提出的统一逻辑规约建立在匹配逻辑的模式和模式匹配语义上，利用了匹配逻辑基于最简核心定义新理论的关键思

想，基于匹配逻辑定义了不同神经网络的形式语义和神经网络的常用性质。该逻辑框架定义了描述神经网络中线性操作、动态传播和时序行为的逻辑结构，不仅可包含针对 ReLU 神经网络的时序逻辑 (ReTL)，还可扩展到对具有不同激活函数的前馈神经网络，现实中常用的网络结构如卷积神经网络和循环神经网络的行为规范。此外，该框架还具有针对新激活函数设计及操作的灵活性和可扩展性。

论文的第三部分介绍了不确定性标准导向的测试用例生成技术。基于抽象的安全分析在一定程度上提高了可信保障技术的可扩展性，然而严格的性质分析和验证往往需要很大的计算强度，无法高效地处理实际应用中的大规模深度学习系统。与之相比，测试技术通过测试用例集触发系统潜在错误行为。尽管无法提供严格的性质保证，但是测试技术的可扩展性强，计算强度低，能够高效地处理复杂深度学习系统，与基于模型的分析 and 验证在系统可信性保障上具有很好的互补性。从探究深度学习系统决策不确定性和对抗攻击鲁棒性缺陷的关系出发，本论文基于确定性执行和贝叶斯执行上的不确定性度量提出了不确定性覆盖标准，进一步利用不确定性指标设计目标函数，提出了由不确定性覆盖标准导向的测试用例生成技术，所生成的测试用例具有更多样的不确定性模式，填补了已有测试和攻击技术未覆盖到的数据模式。

论文的第四部分提出了深度学习系统后门攻击的安全检测和鲁棒性保障技术。上述针对深度学习系统对抗攻击的安全分析技术无法保障系统开发阶段的安全性、可信性，因此研究运行时分析技术以保障系统开发过程中的安全性是可信保障框架中不可或缺的一部分。联邦学习是一类强大的分布式学习范式，通过聚合一组分布式客户端的模型更新来学习一个全局模型。在开发过程中，攻击者可通过控制分布式客户端，向客户端的训练数据中加入后门，进而影响聚合更新后的全局模型鲁棒性。本文提出了一套分布式客户端差异分析框架，可集成到联邦学习程序中作为运行时分析组件。基于采样的差异分析技术，能够有效刻画由后门攻击引起的分布式客户端上的模型偏差，所提出的异常值迭代更新算法可准确检测后门嵌入下的客户端与诚实客户端的行为差异，通过对学习过程中识别到的恶意客户端进行聚合移除，保障全局模型鲁棒性。

本论文研究内容为深度学习系统在对抗攻击和后门攻击下的安全分析和鲁棒性保障技术，涵盖了基于抽象模型的安全分析，形式规约，高效测试和运行时分析技术，构成了一套基于智能系统行为语义，可扩展性强的可信性保障框架，为深度学习系统面临的安全性、可信性问题提供了有效的解决方案，对于推动可信深度学习系统这一新兴领域的发展具有重要意义。

关键词：深度学习系统，自动机提取，逻辑规范，测试用例生成，运行时分析

Towards Trustworthiness Assurance of Deep Learning Systems with Certification Techniques

Xiyue Zhang (Applied Mathematics)

Directed by Prof. Meng Sun

ABSTRACT

In recent years, with the rapid development of computing resources and artificial intelligence related technologies, deep learning (DL) systems have been widely used in a range of applications. Such intelligent systems that incorporate neural networks as the decision-making components have achieved superior performance in prediction accuracy and efficiency when dealing with complex tasks such as computer vision and machine translation. Meanwhile, the deployment of DL systems in safety-critical applications, such as autonomous driving and medical diagnosis, has raised public concerns about the correctness and robustness of these systems. In fact, like traditional hardware and software systems, DL systems suffer from severe safety and trustworthiness issues, including the lack of decision robustness and interpretability of the decision-making process. These issues have hindered more widespread applications of DL systems, especially in safety- and security-critical scenarios.

Certification techniques have demonstrated their usefulness and effectiveness in assuring the safety and reliability of traditional software and hardware systems. However, compared with traditional systems, DL systems have unique internal structure and decision logic. Existing techniques for the trustworthiness assurance of DL systems tend to build on the certification practices for traditional systems and fall short in supporting the decision logic of DL systems. Moreover, existing certification techniques for DL systems are faced with the challenge of scalability, which makes it hard to meet the practical needs for complicated tasks. Therefore, this thesis concentrates on the robustness defects of DL systems sourcing from evasion and backdoor attacks and aims to develop certification techniques dedicated to DL systems based on their unique characteristics. The concrete research contents include abstract model extraction, unifying logical specification for neural networks, uncertainty-guided test case generation, and differential analysis of distributed clients in federated learning.

The first part of the thesis proposes the abstract model extraction algorithm for stateful

DL systems. Existing safety analysis and verification techniques for DL systems are limited to dealing with neural networks of small size and specific types. The complex computation and internal design make it challenging to directly analyze the neural networks. This calls for the need of abstraction techniques with high precision and strong scalability, which could serve as the basis for further analysis and verification of large-scale DL systems. The proposed weighted automata extraction algorithm constructs the state transition model of the target system through abstraction. Further, the extracted weighted automata are used to perform safety analysis of the target system against adversarial attacks. In addition, the extracted weighted automaton can serve as the global interpretation of the target system, which improves the interpretability of the decision process of DL systems. In the extraction algorithm, prediction confidence is used as guidance to build abstract states and feature vectors with semantics similarity are used to extract the transition rules, which greatly enhances the advantage of our approach in extraction precision over the state-of-the-arts.

The second part of the thesis proposes a unifying logical framework for the specification of neural networks. The design and development of a specification framework are critical for safe and secure deployment of intelligent systems in real practice, which are also important prerequisites for the safety analysis of systems. A formal and rigorous logical specification framework can not only characterize behaviors and properties of neural networks, but also serve as the foundation to proof certification and model checking for safety guarantee of neural network behaviors. The unifying logical framework proposed in this thesis builds on the patterns and pattern matching semantics of matching logic and leverages its key insight of axiomatically defining a new specification (also called a theory) based on a simple and minimal core. Generally, we define formal semantics of a variety of neural networks and common properties of neural networks based on matching logic. This specification framework defines generic logical constructs to characterize linear operations, dynamic propagation, and temporal behaviors of neural networks, which not only subsumes ReTL (ReLU Temporal Logic), but also offers good extensibility to neural network variants with different activation functions, as well as realistic neural network architectures, convolutional neural networks and recurrent neural networks. In addition, it has the flexibility of incorporating new activation function designs and operations that will arise in the rapidly developing field of deep learning.

The third part of the thesis introduces uncertainty-guided automated test case generation. Abstraction-based safety analysis and verification improve the scalability of trustworthiness assurance techniques to a certain extent. However, rigorous property analysis and verification

often require intensive computation and fail to deal with large-scale DL systems efficiently in real practice. In contrast, the testing technique identifies and reveals erroneous behaviors of systems through a set of test cases. Although it cannot provide rigorous property guarantees, the testing technique has strong scalability, low computational intensity, and is able to handle complex DL systems efficiently, which lends itself a good complement to the model-based analysis and verification for the trustworthiness assurance of DL systems. Starting from the investigation of the relation between the intrinsic uncertainty nature of deep learning decisions and the robustness defects from adversarial attacks, this thesis proposes the coverage criteria in terms of uncertainty based on the uncertainty metrics from both deterministic and Bayesian executions. Furthermore, we design objective functions and present the automated test case generation approach guided by uncertainty metrics. The generated test cases have more diverse uncertainty patterns, which fill in the gap that is not covered by existing testing and attack techniques.

The fourth part of the thesis proposes the backdoor detection and robustness assurance techniques against backdoor attacks for DL systems. The aforementioned trustworthiness assurance techniques against evasion attacks cannot assure the safety and trustworthiness of DL systems during the development process. Therefore, runtime analysis techniques are an indispensable part of the DL trustworthiness assurance solutions. Federated learning (FL), as a powerful distributed learning paradigm, trains a global model through aggregation on the updates of a set of clients. In the development process, an adversary can control one or more distributed clients and inject a backdoor to the local training data of the clients, thereby affecting the robustness of the global model after update aggregation. In this thesis, we propose a novel differential analysis framework on distributed clients, which can be integrated into the FL procedure as a runtime analysis component. We exploit sampling-based representation differential analysis to capture the model deviation of local clients caused by backdoor attacks. The proposed iterative algorithm for outlieriness quantification can accurately detect the difference between malicious and honest clients, and eliminate the identified misbehaved clients from aggregation for each learning round, which improves the robustness of global models against backdoor attacks.

This thesis makes contributions to the safety, security and robustness analysis of DL systems against evasion and backdoor attacks, which incorporates abstraction-based safety analysis, formal specification, efficient testing and runtime analysis techniques. They constitute a comprehensive framework for the trustworthiness assurance of DL systems with deliberate

consideration of DL behavior semantics and strong scalability. This thesis provides effective solutions to the safety, security and trustworthiness problems faced by DL systems and is of great significance in promoting the development of emerging research on trustworthy DL systems.

KEYWORDS: Deep Learning Systems, Automata Extraction, Logical Specification, Test Case Generation, Runtime Analysis

Contents

Chapter 1 Introduction	1
1.1 Motivation and Challenges	2
1.1.1 Abstract Model Extraction	4
1.1.2 Unifying Logical Framework	5
1.1.3 Uncertainty-Guided Testing	6
1.1.4 Differential Analysis for Runtime Backdoor Detection	7
1.2 Contributions	9
1.3 Thesis Outline	12
Chapter 2 Preliminaries and Related Works	15
2.1 Preliminaries	15
2.1.1 Neural Networks	15
2.1.2 Recurrent Neural Networks	17
2.1.3 Backdoor Attacks	18
2.1.4 Bayesian Uncertainty	19
2.2 Related Work	22
2.2.1 Automata Extraction	22
2.2.2 Attack and Defense	23
2.2.3 Specification for Neural Networks	23
2.2.4 Verification for Neural Networks	24
2.2.5 Testing for Neural Networks	25
2.2.6 Robust Distributed Learning	25
2.2.7 Network Representation Comparison	26
Chapter 3 Weighted Automata Extraction	29
3.1 Weighted Automata Extraction Scheme	29
3.1.1 Weighted Finite Automata	29
3.1.2 Overview of the Extraction Approach	29
3.2 Technical Challenges and Solutions in Extraction	32
3.2.1 Decision-Guided State Abstraction	32
3.2.2 Context-Aware State Composition	34

3.2.3	Synonym Transition for New Data Tolerance	35
3.3	Experiments	36
3.3.1	Datasets and Baselines	36
3.3.2	Experimental Settings	36
3.3.3	Evaluation Metrics	37
3.3.4	RQ1: Approximation Precision	37
3.3.5	RQ2: Precision Enhancement by State Composition	39
3.3.6	RQ3: Synonym Transition Effectiveness	39
3.3.7	RQ4: WFA Extraction from Black-Box RNNs	40
3.3.8	Experiments on GRU Models	42
3.4	Adversarial Example Generation	46
3.5	Summary	48
Chapter 4	Unifying Logical Framework	49
4.1	Matching Logic Basics	49
4.1.1	Matching Logic Syntax	49
4.1.2	Matching Logic Semantics	50
4.1.3	Important Constructs	51
4.2	Unifying Logical Framework for Neural Networks	53
4.2.1	Defining Linear Operations	53
4.2.2	Defining Dynamic Propagation	55
4.2.3	Defining Atomic Formulas and Temporal Operations	57
4.3	Instance: Feed-Forward Neural Networks	57
4.3.1	ReTL: A Logic for ReLU Networks	58
4.3.2	Defining ReTL in Matching Logic	60
4.3.3	Defining Standard Models of Feed-Forward ReLU Networks	61
4.3.4	Instances of Neural Networks Using Other Activation Functions	66
4.4	Instance: Convolutional Neural Networks	67
4.4.1	Convolution and Pooling Operations	67
4.4.2	Propagation of Convolutional Neural Networks	68
4.4.3	Defining Standard Models of Convolutional Neural Networks	70
4.5	Instance: Recurrent Neural Networks	70
4.5.1	Propagation of Recurrent Neural Networks	71
4.5.2	Defining Standard Models of Recurrent Neural Networks	72

4.6	Specifying Neural Network Properties	73
4.6.1	Robustness	73
4.6.2	Interval Property	73
4.6.3	Monotonicity	74
4.6.4	Fairness	74
4.7	Summary	75
Chapter 5 Test Generation Guided by Uncertainty		77
5.1	Research Question Formulation	77
5.2	Empirical Study	79
5.2.1	Subject Datasets and Data Preparation	80
5.2.2	RQ1: Empirical Study on Uncertainty Metrics	81
5.2.3	Characterizing Data Behavior	83
5.2.4	RQ2: Categorization of Existing Data	83
5.3	Test Generation	85
5.4	Evaluation	88
5.4.1	RQ3: Usefulness of Test Data Generation	88
5.4.2	RQ4: Evaluation on Defense Techniques	91
5.5	Discussion about Threat to Validity	94
5.6	Summary	94
Chapter 6 A Differential Analysis Framework for Runtime Backdoor Detection		97
6.1	Attack Model	97
6.2	Differential Analysis Framework	98
6.2.1	Framework Design	98
6.2.2	RDM Extraction	99
6.2.3	Client Differential Analysis	101
6.2.4	Iterative Local Outlier Factor Update	101
6.3	Experiments	103
6.3.1	Experiment Configuration	103
6.3.2	Experiment Results on Detection Accuracy	105
6.3.3	Deployment as Runtime Detection Component	107
6.3.4	Runtime Efficiency	112
6.4	Effectiveness of Iterative Algorithm	113
6.5	Client Dissimilarity Demonstration	115

6.6 Summary	117
Chapter 7 Conclusion and Future Work	119
7.1 Conclusion	119
7.2 Future Work	121
Bibliography	123
攻读博士期间发表的论文及其他成果	137
致谢	141
北京大学学位论文原创性声明和使用授权说明	143

攻读博士期间发表的论文及其他成果

个人简介

张喜悦，女，2013年9月至2017年6月就读于北京大学数学科学学院，2017年6月获得信息与计算科学学士学位。2017年9月至2022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. **Xiyue Zhang**, Xiaoning Du, Xiaofei Xie, Lei Ma, Yang Liu and Meng Sun. Decision-Guided Weighted Automata Extraction from Recurrent Neural Networks. in Proceedings of AAAI 2021, pages 11699-11707, 2021.
2. **Xiyue Zhang**, Xiaofei Xie, Lei Ma, Xiaoning Du, Qiang Hu, Yang Liu, Jianjun Zhao and Meng Sun. Towards Characterizing Adversarial Defects of Deep Learning Software from the Lens of Uncertainty. in Proceedings of ICSE 2020, pages 739-751, ACM, 2020.
3. **Xiyue Zhang**. Uncertainty-Guided Testing and Robustness Enhancement for Deep Learning Systems. in Proceedings of ICSE 2020 (Companion Volume), pages 101-103, ACM, 2020.
4. **Xiyue Zhang**, Yi Li and Meng Sun. Towards a Formally Verified EVM in Production Environment. in Proceedings of COORDINATION 2020, LNCS 12134, pages 341-349, Springer, 2020.
5. **Xiyue Zhang**, Weijiang Hong, Yi Li and Meng Sun. Reasoning about Connectors Using Coq and Z3. *Science of Computer Programming*, vol. 170, pages 27-44, 2019.
6. **Xiyue Zhang** and Meng Sun. SMT-based Modeling and Verification of Cloud Applications. in Proceedings of SERVICES 2019, LNCS 11517, pages 1-15, Springer, 2019.
7. **Xiyue Zhang**, Yi Li, Weijiang Hong and Meng Sun. Using Recurrent Neural Network to Predict Tactics for Proving Component Connector Properties in Coq. in Proceedings of TASE 2019, pages 107-112, IEEE, 2019.
8. **Xiyue Zhang** and Meng Sun. Towards Formal Modeling and Verification of Probabilistic Connectors in Coq. in Proceedings of SEKE 2018, pages 385-390, KSI Research

- Inc. and Knowledge Systems Institute, 2018.
9. **Xiyue Zhang**. Modeling and Verification of Component Connectors. in Proceedings of ICFEM 2018, LNCS 11232, pages 419-422, Springer, 2018.
 10. **Xiyue Zhang**, Weijiang Hong, Yi Li and Meng Sun. Reasoning about Connectors in Coq. in Proceedings of FACS 2016, pages 172-190, LNCS 10231, Springer, 2017.
 11. Weidi Sun, Yuteng Lu, **Xiyue Zhang** and Meng Sun. DeepGlobal: a Framework for Global Robustness Verification of Feedforward Neural Networks. *Journal of Systems Architecture*, 2022.
 12. Xiaokun Luan, **Xiyue Zhang** and Meng Sun. Using LSTM to Predict Tactics in Coq. in Proceedings of SEKE 2021, pages 132-137, KSI Research Inc. and Knowledge Systems Institute, 2021.
 13. Weidi Sun, Yuteng Lu, **Xiyue Zhang** and Meng Sun. DeepGlobal: a Global Robustness Verifiable FNN Framework. in Proceedings of SETTA 2021, LNCS 13071, pages 22-39, Springer, 2021.
 14. Yi Li, **Xiyue Zhang**, Yuanyi Ji and Meng Sun. A Formal Framework Capturing Real-Time and Stochastic Behavior in Connectors. *Science of Computer Programming*, vol. 177, pages 21-40, 2019.
 15. Bai Xue, Yang Liu, Lei Ma, **Xiyue Zhang**, Meng Sun and Xiaofei Xie. Safe Inputs Generation for Black-box Systems. in Proceedings of ICECCS 2019, pages 180-189, IEEE, 2019.
 16. Meng Sun and **Xiyue Zhang**. A Relational Model for Probabilistic Connectors based on Timed Data Distribution Streams. in Proceedings of FORMATS 2018, LNCS 11022, pages 125-141, Springer, 2018.
 17. Weijiang Hong, Saqib Nawaz, **Xiyue Zhang**, Yi Li and Meng Sun. Using Coq for Formal Modeling and Verification of Timed Connectors. in Software Engineering and Formal Methods: SEFM 2017 Collocated Workshops, LNCS 10729, pages 558-573, Springer, 2018.
 18. Yi Li, **Xiyue Zhang**, Yuanyi Ji and Meng Sun. Capturing Stochastic and Real-time Behavior in Reo Connectors. in Proceedings of SBMF 2017, pages 287-304, LNCS 10623, Springer, 2017 (**Best Paper Award**).

已投稿论文

1. **Xiyue Zhang**, Xiaohong Chen and Meng Sun. Towards a Unifying Logical Framework for Neural Networks. 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.
3. Zeming Wei, **Xiyue Zhang** and Meng Sun. Extracting Weighted Finite Automata from Recurrent Neural Networks for Natural Languages. Manuscripts.

专利

一种生成黑盒循环神经网络对抗样本的方法，专利号：202111339035.4，张喜悦，孙猛，北京大学（已受理）

软件著作权

循环神经网络的加权有穷自动机提取软件 V1.0，2021SR1997328，北京大学，第 1 完成人

获奖情况

1. 2021 年 10 月. 国家奖学金
2. 2020 年 11 月. 中国软件大会 2020 优秀博士生
3. 2020 年 10 月. 国家奖学金
4. 2020 年 6 月. 北京大学校长奖学金
5. 2019 年 10 月. 廖凯原奖学金
6. 2019 年 6 月. 北京大学校长奖学金
7. 2018 年 10 月. 五四奖学金
8. 2018 年 6 月. 北京大学校长奖学金
9. 2017 年 11 月. SBMF 2017 最佳论文奖
10. 2017 年 9 月. 北京大学校长奖学金

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
4. 2018-2021, 信息物理系统中复杂并发行为的形式化建模与验证
5. 2016-2020, 大规模概率并发实时系统的模型检验
6. 2018-2019, 智能合约安全性建模与验证

致谢

光阴荏苒，转眼间已经来到了博士研究生的最后一个年头。回首博士生阶段时光，心中倍感充实，谨借此机会，向成长发展道路上给予我指导、帮助和鼓励的老师、同学、亲友表达感恩之心，致以最真挚的谢意。

首先，我要由衷地感谢我的导师孙猛教授在博士期间对我的培养、指导和帮助，让我从科研新手逐渐蜕变成了一个更成熟的科研工作者。与孙老师的认识，起始于本科生科研，从第一个课题项目开始，孙老师的耐心教诲和培养，让我从一个科学研究的门外汉，进入了软件可靠性研究的大门，让我感受到了专业课程学习与本科生科研相互促进的成就感，让我学会了工作项目上的朋辈合作，让我学会了如何撰写和分享科学研究成果。在孙老师的指导下，从本科生科研开始的一段段科研实践经历锻炼了我，让我收获了专业知识和能力的提升，更收获了科学研究的精神和方法，这些将让我受益终生。

其次，我要感谢在课题项目上指导、帮助过我，同我深入交流讨论课题方向的老师。在新加坡南洋理工大学交流学习期间，得到了刘杨老师，马雷老师，赵建军老师，谢肖飞博士，杜晓宁博士的指导和帮助，让我顺利地展开了博士论文的研究课题。在与指导老师的交流和讨论中，我学习到了课题项目的研究现状，扩充了相关专业知识，丰富了项目开发经验，进一步学习到了科学研究的方法，提高了独立研究能力，他们的指导与本论文的顺利完成密不可分。感谢在新加坡科技设计大学暑期交流期间指导我展开智能合约验证研究的孙军老师。感谢中国科学院软件研究所的薛白老师同我讨论合作黑盒系统安全性的研究工作，拓宽了我的知识视野。在课题研究的开展过程中，指导老师们对科学研究的深刻见解、创新想法给我留下了深刻印象，为我之后的科学研究提供了偌大的启发和帮助。

我还要感谢教授我专业知识以及给予我课题指导的老师，感谢在北大教授我博士生专业课程知识的曹永知老师，陈向群老师，金芝老师，林作铨老师，马尽文老师，王捍贫老师，夏壁灿老师，徐茂智老师，杨建生老师。感谢在中科院软件所开设的形式化方法暑期学校为我授课的 Prof. Martin Fränzle, Prof. Holger Hermanns, Prof. Joost-Pieter Katoen, Prof. Stefan Mitsch, Prof. Bow-Yaw Wang. 感谢曾在会议及讨论班上同我展开讨论交流的白光冬老师，卜磊老师，陈振邦老师，贺飞老师，姜宇老师，李建文老师，刘滄老师，刘万伟老师，刘志明老师，蒲戈光老师，裘宗燕老师，余志坤老师，施志平老师，宋富老师，王戟老师，熊英飞老师，许智武老师，詹博华老师，詹乃军老师，张立军老师，张民老师，张敏老师，赵永望老师，Prof. Bernhard Aichernig, Prof. Farhad

Arbab, Prof. Hamideh Afsarmanesh, Prof. Luis Barbosa, Prof. David Jansen, Dr. Renato Neves, Prof. Ernst-Rüdiger Olderog, Prof. Anders P. Ravn 等各位老师。

此外，我还要感谢在项目研究中同我一起合作、攻坚克难的同学、朋友们。感谢李屹博士在协调系统验证等一系列科研工作中的指导和帮助，帮我度过了科研初期的迷茫，顺利展开了软件可靠性验证的研究。感谢洪伟疆同学在我最初的科研项目中提供的帮助，同我一起从零出发了解和探索一个全新领域的知识，共同成长、共同进步。感谢陈霄泓博士在神经网络逻辑规范课题中的合作和帮助，推进了逻辑规范框架的顺利开展。感谢冀元祎在随机协调系统建模中的合作和帮助，让我提升了概率行为建模的知识和经验。感谢 Dr. Saqib Nawaz 在实时连接件建模工作中的探讨与合作，促进了协调系统实时行为扩展的顺利展开。感谢孙纬地、卢煜腾同学在深度学习鲁棒性上的合作与探讨，让我拓展了课题知识与经验。感谢薛骁勇同学在分布式学习安全上的合作与探讨，促进了课题研究和评估的顺利推进。感谢栾晓坤同学在定理证明策略预测上的合作与讨论，让我进一步收获了在自动定理证明策略合成方向的知识与经验。感谢胡强同学在深度学习防御技术上的合作，促进了测试用例评估的顺利展开。

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。学习科研上，除了同我展开课题合作的同学，还有同我一起在讨论班上学习交流的刘艾，王顺，徐鹤元，张琦，冯逸群，卜昊，杨晓宇，李忆。在北大攻读博士的时光里，我还要感谢同我一起走过学习、科研和生活上的点点滴滴的罗姗姗同学，伴我走过博士初始阶段科研学习和生活中酸甜苦辣的陈琪同学，同我畅聊课题走向、毕业事宜、工作申请的李昊坤同学。在新加坡南洋理工大学交流期间，除了同我展开课题合作的同学、朋友，周远博士，唐文兵，雷予斯，范玲玲博士，陈森博士，吴修衡，刘晔，刘承威，陈康杰，孙剑文博士，陈泓旭博士，唐赟，姜柯，朱宸光，David Berend 在工作和生活中给予了我很大的慰藉，在科研学习之外增添了生活中的欢乐。在暑期新加坡科技设计大学交流期间，王竟亦博士，陈宇奇博士，高博帮助我很好地适应了新的工作环境，在博士阶段即将展开之际给予了我很大的鼓励以及课题选择上的启发。

最后，我衷心地感谢我的父母、家人，感谢他们对我一路上的鼓励和支持，感谢他们无私的奉献和关怀，让我得以顺利完成学业。感谢赵健先生对我科研、学习和生活上的鼓励和帮助，同我一起走过万水千山，让我得以充满信心和勇气去迎接工作和生活中的挑战。

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；
学校有权保留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

