北京大学

博士研究生学位论文

题目：基于测试的机器学习系统安全加强研究

姓　　名：卢煜腾

学　　号：2001110047

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 三 月

版权声明

摘要

随着人工智能的快速发展和普及，机器学习技术已被广泛应用于模式识别、语音识别、机器翻译等多个领域，取得了显著的进展和突破，其表现甚至可以超越人类，为人类生活带来了巨大的便利。特别值得注意的是，机器学习技术在自动驾驶、人脸识别、医疗诊断等安全攸关领域扮演着越来越重要的角色。然而，随着以对抗样本为代表的隐藏风险被揭示以及一系列令人痛心的事故在安全攸关领域频繁发生，人们开始质疑机器学习技术是否具有足够的鲁棒性和安全性来保障其在安全攸关领域的应用。机器学习系统决策过程的不可解释性使得这一问题一直未能得到很好解决。因此，本文利用变异测试、模糊测试等技术，对机器学习系统的安全性加强方法进行研究，以期为机器学习技术提供更加坚实的安全保障。

变异测试是一种评估测试数据质量和识别系统缺陷的成熟技术；模糊测试则是通过构造和执行测试用例来识别系统潜在漏洞的测试技术。这些技术基于传统软件的体系结构和决策逻辑，在传统软件质量保障方面已经取得了巨大的成功。然而，由于机器学习系统与传统软件系统在行为和决策逻辑方面存在本质区别，因此传统软件领域的安全保障技术并不能直接适用于机器学习系统。

本文涵盖三类重要的机器学习范式：监督学习、无监督学习和强化学习。针对监督学习的特性，本文对模糊测试和变异测试技术进行了优化和集成，基于这一集成测试框架提出了一套用于诊断和修复异常神经网络的方法，并将提出的技术与方法应用于监督学习系统。针对无监督学习和强化学习的特性，本文分别设计并实现了相应的变异测试方法，在实验中取得了很好的效果。

本文的第一部分针对监督学习系统，基于强化学习对变异测试和模糊测试技术进行优化，并将优化得到的技术进行集成，提出了一个理论框架 *RGChaser*，用于生成高质量的测试用例，通过变异测试自动生成符合预先定义目标的变异体，通过模糊测试自动处理多样性目标并以较低开销生成测试用例。实验结果表明，与已有研究相比，这一框架在生成目标测试用例和变异体方面均具有更高的效率和成功率。论文还提供了一个基于 *RGChaser* 开发的开源 GUI 工具 RGChaser，通过在实验中使用该工具分析神经网络中潜在问题的根本原因，对其应用能力进行了验证。此外，论文提出了一套基于测试对异常神经网络中的问题进行自动诊断并给出修复建议的方法 *MRepair*，可以处理神经网络中损失震荡、收敛缓慢等各种常见问题。

本文的第二部分提出了一种针对无监督学习系统的变异测试方法 *MTUL*，在数据和算法层面分别构建了相应的变异框架。对于聚类分析、生成对抗网络等无监督学习

技术，该方法从数据和算法层面中的多个视角提出了一系列变异算子，定义了相应的变异分数，并将 *MTUL* 与自编码器相结合，提出了一种构造对抗样本的新方法。*MTUL* 不仅能用于筛选高质量的测试数据，还能评估系统的稳定性和抵御风险的能力。论文开发了一个嵌入了针对生成对抗网络的变异测试技术的开源 GUI 工具 MTGAN，以便开发人员对生成对抗网络进行评估。

本文的第三部分提出了针对强化学习系统的变异测试方法。由于强化学习基于智能体与环境的交互，对强化学习系统无法提供类似于监督学习系统或无监督学习系统中的测试数据。该方法定义了一系列以元素级别算子和智能体级别算子为主的变异算子，用于模拟强化学习系统可能面临的问题。本文从多个角度（例如探索-利用困境）考虑变异算子的设计，以期充分覆盖潜在错误类型，提高变异算子集合的完备性，根据强化学习的特点设计了专用的变异分数和变异测试框架。这一方法有望用于指导强化学习系统测试环境的构建，揭示强化学习系统的潜在错误，并协助系统设计者构建更符合预期性能的强化学习系统。

本论文的研究内容为面向机器学习系统安全保障的测试技术，涵盖了监督学习、无监督学习和强化学习，优化了针对监督学习系统的变异测试方法和模糊测试方法，基于此构造了一套可信性保障框架，为无监督学习系统和强化学习系统设计了专用的变异测试方法，为机器学习系统的可信性、安全性问题提供了高效的解决方案，对机器学习系统在安全攸关领域的应用及可信机器学习技术的发展有重要意义。

关键词：机器学习系统，监督学习，无监督学习，强化学习，变异测试，模糊测试

Testing-based Machine Learning Systems Safety Strengthening

Yuteng Lu (Applied Mathematics)
Directed by Prof. Meng Sun

ABSTRACT

With the rapid development and popularization of artificial intelligence, machine learning technique has been widely applied in various fields, such as pattern recognition, speech recognition, machine translation, and more, achieving significant progress and breakthroughs. Its performance can even surpass that of humans, bringing great convenience to human life. It is particularly noteworthy that machine learning technique plays an increasingly important role in safety-critical domains, including autonomous driving, facial recognition, and medical diagnosis. However, with the exposure of hidden risks represented by adversarial examples and a series of heartbreaking accidents occurring frequently in safety-critical domains, people have begun to question the ability of machine learning technique to provide the necessary levels of robustness and security for its application in these domains. The lack of interpretability in the decision-making process of machine learning systems has made this problem persistently unresolved. Therefore, this thesis investigates methods to strengthen the safety of machine learning systems using testing techniques such as mutation testing and fuzz testing, aiming to provide a more solid trustworthiness guarantee for machine learning technique.

Mutation testing has been proven to be a mature technique for evaluating the quality of testing data and identifying system defects; while fuzz testing is a testing technique for identifying potential system vulnerabilities by constructing and executing test cases. These techniques, based on the architecture and decision logic of traditional software, have achieved great success in traditional software quality assurance. However, due to the fundamental differences in behavior and decision-making logic between machine learning systems and traditional software systems, safety assurance techniques from the traditional software domain cannot be directly applied to machine learning systems.

This thesis covers three important machine learning paradigms: supervised learning, unsupervised learning, and reinforcement learning. According to the characteristics of supervised learning, this thesis optimizes and integrates the techniques of fuzz testing and mutation testing, proposes an approach for diagnosing and repairing abnormal neural networks based on the integrated testing framework, and applies the proposed techniques and approach to supervised learning systems. In allusion to the characteristics of unsupervised learning and reinforcement learning, this thesis has designed and implemented corresponding mutation testing approaches, which have achieved promising results in experiments.

The first part of this thesis focuses on supervised learning systems, optimizes mutation testing and fuzz testing techniques based on reinforcement learning, and integrates the optimized techniques into a theoretical framework *RGChaser* to generate high-quality test cases. *RGChaser* automates the generation of mutants that meet pre-defined targets through mutation testing, and produces test cases with low overhead by automatically handling diverse targets through fuzzing. The experimental results demonstrate that the proposed framework has higher efficiency and success rates in generating target test cases and mutants than existing approaches. The thesis also provides an open-source GUI tool RGCHASER developed based on *RGChaser*. The application capability of the tool have been validated in experiments analyzing the root causes of potential problems in neural networks. Furthermore, the thesis proposes a testing-based approach *MRepair* for automatically diagnosing and providing repair suggestions for problems in abnormal neural networks, such as Oscillating Loss and Slow Convergence.

The second part of this thesis proposes a mutation testing approach *MTUL* for unsupervised learning systems, which builds corresponding mutation frameworks at both the data and algorithm levels. This thesis proposes a series of mutation operators for unsupervised learning techniques, such as cluster analysis and generative adversarial networks (GANs), from multiple perspectives at the data and algorithm levels, and defines the corresponding mutation scores. Additionally, a new approach for constructing adversarial examples is developed by combining *MTUL* with autoencoders. *MTUL* can be used not only to screen high-quality test data, but also to aid in evaluating the stability and risk-resistance capabilities of systems. The thesis develops an open-source GUI tool, MTGAN, that incorporates the mutation testing technique for GANs, enabling developers to evaluate GANs efficiently.

The third part of this thesis proposes a mutation testing approach for reinforcement learning systems. As reinforcement learning is built on the interaction between an agent and its environment, it is not possible to provide test data for reinforcement learning systems similar to those for supervised or unsupervised learning systems. The approach defines a series of mutation operators, primarily consisting of element-level and agent-level operators, which are used to simulate the issues that reinforcement learning systems may encounter. This thesis addresses the design of mutation operators from various perspectives (*e.g.* the exploration-exploitation dilemma), in order to comprehensively cover potential error types and enhance the overall completeness of the mutation operator set. Furthermore, the thesis designs dedicated mutation scores and mutation testing frameworks tailored to the unique characteristics of reinforcement learning. The presented approach is expected to guide the construction of test environments for reinforcement learning systems, reveal potential errors in such systems, and assist system designers in building reinforcement learning systems that better meet expected performance.

The research content of this thesis is focused on testing techniques for safety assurance of machine learning systems, covering supervised learning, unsupervised learning, and reinforcement learning. A framework for safety assurance of supervised learning systems is constructed based on the optimization and integration of mutation testing and fuzz testing techniques. Dedicated mutation testing approaches are designed for unsupervised learning systems and reinforcement learning systems. The thesis provides efficient solutions to the reliability and safety issues faced by machine learning systems, which is of great significance to the application of machine learning systems in safety-critical fields and the development of trusted machine learning technique.

KEY WORDS: Machine Learning systems, Supervised Learning, Unsupervised Learning, Reinforcement Learning, Mutation Testing, Fuzz Testing

Contents

Chapter 1 Introduction and Motivation 1

1.1 Motivation and Challenges 1

1.1.1 For Supervised Learning Systems 8

1.1.2 For Unsupervised Learning Systems 9

1.1.3 For Reinforcement Learning Systems 10

1.2 Contributions 11

1.3 Thesis Outline 14

Chapter 2 Preliminaries and Related Works 15

2.1 Preliminaries 15

2.1.1 Machine Learning 15

2.1.2 Mutation Testing 18

2.1.3 Fuzz Testing 19

2.1.4 Security Vulnerabilities of ML Systems 20

2.1.5 Automated Program Repair 21

2.2 Related Works 22

2.2.1 Security Issues within Machine Learning 22

2.2.2 Testing for ML systems 24

2.2.3 Mutation Testing 27

2.2.4 Fuzz Testing 30

2.2.5 Neural Network Repair 33

Chapter 3 RGChaser: A RL-guided Fuzz and Mutation Testing Framework for Supervised Learning Systems 35

3.1 Overview 35

3.2 RL-guided Fuzz Testing 37

3.2.1 Elements 37

3.2.2 RL-guided Fuzzing Process 39

3.3 RL-guided Mutation Testing 41

3.3.1 Elements 41

3.3.2 RL-guided Mutation Process 42

3.4 Framework for Efficient Screening of Test Data 44
3.5 Mutation Operators for Diverse Mutants 46
3.5.1 Mutation Operators for CNNs 46
3.5.2 Mutation Operators for Continuous Mutation Spaces.........................47
3.6 RGChaser: A Learning-Based GUI Tool..47
3.7 Finding Roots of DNN Problems with RGChaser 51
3.8 *MRepair*: Towards Repairing Abnormal Neural Networks 52
3.9 Evaluation 55
3.9.1 Evaluation Settings 56
3.9.2 Research Question 1 57
3.9.3 Research Question 2 61
3.9.4 Research Question 3 63
3.9.5 Research Question 4 64
3.9.6 Research Question 5 65
3.10 Summary 66
Chapter 4 Mutation Testing of Unsupervised Learning Systems.......................69
4.1 Overview 69
4.2 Mutation Operators and Scores for Unsupervised Learning Systems 71
4.2.1 Mutation Operators..71
4.2.2 Mutation Scores 75
4.3 *MUAE*: The Combination of *MTUL* and Autoencoder...............................79
4.4 MTGAN: A GUI Tool Integrating Mutation Testing Technique for GANs 82
4.5 Evaluation 83
4.5.1 Datasets 84
4.5.2 Research Question 1 84
4.5.3 Research Question 2 86
4.5.4 Research Question 3 87
4.6 Summary 89
Chapter 5 Mutation Testing Specific for Reinforcement Learning Systems...........91
5.1 Overview 91
5.2 Mutation Testing Specific to RL Systems 92
5.2.1 Element-level Mutation Operators...92
5.2.2 Agent-level Mutation Operators 95

5.2.3 Mutation Testing Framework for RL Systems 96
5.3 Comprehensive Mutation 96
5.3.1 Reward Rules 97
5.3.2 Exploration and Exploitation 99
5.3.3 Experience replay 100
5.3.4 Continuous State Spaces 100
5.4 Mutation Scores Specific to RL Systems 101
5.5 Design of Test Environments 103
5.6 Evaluation 107
5.6.1 Research Question 1 110
5.6.2 Research Question 2 111
5.7 Assist in Constructing Better Systems 113
5.7.1 For Robustness 113
5.7.2 For Reward Rules 114
5.8 Summary 115

Chapter 6 Conclusion and Future Work 117
6.1 Conclusion 117
6.2 Future Work 122

Bibliography 125

攻读博士期间发表的论文成果及其他情况 139

致谢 143

北京大学学位论文原创性声明和使用授权说明 145

攻读博士期间发表的论文成果及其他情况

个人简历

卢煜腾，男，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 月获理学学士学位。2017 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. **Yuteng Lu**, Weidi Sun, and Meng Sun. Towards Mutation Testing of Reinforcement Learning Systems. Journal of Systems Architecture, vol. 131, 102701, 2022.
2. **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, pages 12-23, IEEE, 2022.
3. **Yuteng Lu**, Kaicheng Shao, Weidi Sun, and Meng Sun. MTUL: Towards Mutation Testing of Unsupervised Learning Systems. in Proceedings of SETTA 2022, LNCS 13649, pages 22-40, Springer, 2022.
4. **Yuteng Lu**, Weidi Sun and Meng Sun. Mutation Testing of Reinforcement Learning Systems. in Proceedings of SETTA 2021, LNCS 13071, pages 143-160, Springer, 2021.
5. **Yuteng Lu**, Weidi Sun, Guangdong Bai, Meng Sun. DeepAuto: A First Step Towards Formal Verification of Deep Learning Systems. in Proceedings of SEKE 2021, pages 172-176, KSI Research Inc. and Knowledge Systems Institute, 2021.
6. **Yuteng Lu** and Meng Sun. Modeling and Verification of IEEE 802.11i Security Protocol in UPPAAL for Internet of Things. International Journal on Software Engineering and Knowledge Engineering, vol. 28, no. 11n12, pages 1619-1636, 2018.
7. **Yuteng Lu** and Meng Sun. Modeling and Verification of IEEE 802.11i Security Protocol for Internet of Things. In Proceedings of SEKE 2018, pages 270-275, KSI Research Inc. and Knowledge Systems Institute, 2018.
8. Weidi Sun, **Yuteng Lu**, Xiyue Zhang, and Meng Sun. DeepGlobal: a Framework for Global Robustness Verification of Feedforward Neural Networks. Journal of Systems Architecture, vol. 128, 102582, 2022.
9. Meng Sun, **Yuteng Lu**, Yi-chun Feng, Qi Zhang and Shaoying Liu. Modeling and Verifying the CKB Blockchain Consensus Protocol. Mathematics, vol. 9(22), 2954, 2021.

10. Yi-Chun Feng, **Yuteng Lu**, Meng Sun. Modeling and Verification of CKB Consensus Protocol in UPPAAL. in Proceedings of SEKE 2021, pages 150-153, KSI Research Inc. and Knowledge Systems Institute, 2021.
11. Weidi Sun, **Yuteng Lu**, Meng Sun. Are Coverage Criteria Meaningful Metrics for DNNs? in Proceedings of IJCNN 2021, pages 1-8, IEEE, 2021.
12. 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.
13. Qi Zhang, **Yuteng Lu** and Meng Sun. Modeling and Verification of the Nervos CKB Block Synchronization Protocol in UPPAAL. in Proceedings of BlockSys 2020, CCIS 1267, pages 3-17 , Springer, 2020.
14. Weidi Sun, Xiaoyong Xue, **Yuteng Lu** and Meng Sun. HashC: Making DNNs' Coverage Testing Finer and Faster. in Proceedings of SETTA 2022, LNCS 13649, pages 3-21, Springer, 2022.

已投稿论文

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

获奖情况

1. 2022 年 12 月获五四奖学金
2. 2022 年 12 月获北京大学优秀学生干部
3. 2021 年 12 月获秦宛顺靳云汇奖学金
4. 2021 年 12 月获学术创新奖
5. 2019 年 12 月获北京大学研究生专项学业奖学金
6. 2019 年 12 月获北京大学三好学生
7. 2018 年 12 月获国家奖学金
8. 2018 年 12 月获北京大学三好学生

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. 2018-2019，智能合约安全性建模与验证

致谢

我仍记得 2013 年的 8 月盛夏，走进北大东门后那映入眼帘的图书馆与博雅塔。这书香浸染的地标是我魂牵梦萦的地方，十年前的我期许她引领我的脚步、承载我的追求，转眼十载，她给予了我甘甜，也带我感受过苦涩，亦将成为我的回忆与继续前行的见证。感谢这接纳我的园子，给我一方净土，容许我恣意选择人生。

回顾十年的求学历程，首先要由衷地感谢我的导师孙猛教授。大三时我选修了孙老师的数据结构与算法课程，随后在老师的指导下完成了毕业论文。那时的我尚不了解论文写作的基本流程与规范，对领域内的研究方法也缺乏认知，文章的写作总是略显不成熟。此后，从本科毕业论文到一篇篇会议论文，再至本文成稿，我已经记不清有多少次老师加班加点为我修改文章，通过电话、面谈等各种方式询问我的动向，开解我的困惑。我在自责惭愧的同时，一次又一次地被老师的认真负责深深感动。除却学术道路上的指导，最宝贵的当属孙老师在潜移默化中教导我的为人处世之道。我的导师可以说是近乎无条件地支持着我的人生选择，他会耐心地聆听我的规划，并为我出谋划策，引导我开拓思路，鼓励我大胆追寻。我们的聊天交谈经常长达数小时，甚至会略过午饭时间，我总是舍不得结束每场交谈。于孙老师处，我真正体会到了何为“听君一席话，胜读十年书”。

我还要将感谢致以博士期间教授我专业课程的诸位老师：金芝老师、林作铨老师、马尽文老师、牟克典老师、麻志毅老师、王捍贫老师、谢冰老师、夏壁灿老师、许进老师、徐茂智老师、熊英飞老师、杨建生老师。在向各位老师学习的过程中，我进一步夯实了专业知识，为深入进行学术研究打好了基础。

同时，我还要感谢在项目上给予过我帮助，在讨论时给过我指导和点评的各位老师：白光冬老师、鄂维南老师、刘万伟老师、马尽文老师、蒲戈光老师、裘宗燕老师、佘志坤老师、宋富老师、王戟老师、夏壁灿老师、熊英飞老师、詹乃军老师、张立军老师。老师们的帮助、指导和点评逐步加深了我对于领域内科研方法和学术成果的理解。同时，我也深深敬佩于老师们的学术精神。

进入师门已有八年之久，我要在此感谢每一位并肩奋斗过的师兄、师姐、师弟、师妹：李屹、M. Saqib Nawaz Khan、王译梧、刘艾、徐鹤元、王顺、张喜悦、孙纬地、张琦、冯逸群、杨晓宇、薛骁勇、李忆、卜昊、栾晓坤、邵凯诚。感谢各位在我迷茫懵懂时的倾力相助，无论是深度的合作，还是日常的交流，你们与我的探讨都让我受益良多。山水有来路，早晚复相逢。希望这份情谊能无限延续，来日相会时我们仍是亲密的伙伴。

当然，我也要感谢这个伟大的时代。国泰民安的社会环境让我们能够更踏实地工作学习；包容开发的科学环境给我们带来了以 5G 网络和人工智能为代表的新技术，大大提高了我们的工作效率；欣欣向荣的经济环境也让我们有更多的机会去施展抱负。这一切都来之不易，我应该倍加珍惜。

最后，我要感谢我的父母与妻子，感谢你们成长道路上的支持与陪伴。你们陪我翻越秦岭的山，直面北京的风，你们见证了不同阶段的我。博士数载其中定有诸多心酸与苦难，而你们就是我坚实的靠山、避风的港湾，让我在低谷时也能够迸发出巨大的能量。

行笔至此，也终将要与北大告别。此刻，我再次想起十年前的那个盛夏。惟愿能保留那份赤诚

与初心，行而不辍。

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