IJTCS-FAW: Machine Learning and Formal Method

2:00-5:00, Aug 19, 2022

2:00 - 2:10:

Opening Addressing *(Sun Jun, Singapore Management University)*

2:10 - 2:50:

Invited Talk 1: **QVIP: An ILP-based Formal Verification Approach for Quantized Neural Networks**, published at ASE 2022 *(Yedi Zhang, ShanghaiTech University)*

报告摘要：Deep learning has become a promising programming paradigm in software development, owing to its surprising performance in solving many challenging tasks. Deep neural networks (DNNs) are increasingly being deployed in practice, but are limited on resource-constrained devices owing to their demand for computational power. Quantization has emerged as a promising technique to reduce the size of DNNs with comparable accuracy as their floating-point numbered counterparts. The resulting quantized neural networks (QNNs) can be implemented energy-efficiently. Similar to their floating-point numbered counterparts, quality assurance techniques for QNNs, such as testing and formal verification, are essential but are currently less explored. In this work, we propose a novel and efficient formal verification approach for QNNs. In particular, we are the first to propose an encoding that reduces the verification problem of QNNs into the solving of integer linear constraints, which can be solved using off-of-the-shelf solvers. Our encoding is both sound and complete. We demonstrate the application of our approach on local robustness verification and maximum robustness radius computation. We implement our approach in a prototype tool QVIP and conduct a thorough evaluation. Experimental results on QNNs with different quantization bits confirm the effectiveness and efficiency of our approach, e.g., our approach is two orders of magnitude faster and able to solve more verification tasks in the same time limit than the state-of-the-art methods.

个人介绍：Yedi Zhang is a sixth year Ph.D candidate in ShanghaiTech University, advised by Prof. Fu Song. She received her B.E. degree from Beijing University of Posts and Telecommunications. Her current research interests are automated verification techniques for artificial intelligent systems. Her research has been published in journals and conferences such as Journal of Software, IEEE Access, AAAI, CAV and ASE.

2:50 - 3:30:

Invited Talk 2: **Copy, Right? A Testing Framework for Copyright Protection of Deep Learning Models**, published at S&P 2022 *(Jingyi Wang, Zhejiang University)*

报告摘要：Deep learning models, especially those large-scale and high-performance ones, can be very costly to train, demanding a considerable amount of data and computational resources. As a result, deep learning models have become one of the most valuable assets in modern artificial intelligence. Unauthorized duplication or reproduction of deep learning models can lead to copyright infringement and cause huge economic losses to model owners, calling for effective copyright protection techniques. In this talk, I will present a novel testing framework for deep learning copyright protection: DeepJudge. DeepJudge quantitatively tests the similarities between two deep learning models: a victim model and a suspect model. It leverages a diverse set of testing metrics and efficient test case generation algorithms to produce a chain of supporting evidence to help determine whether a suspect model is a copy of the victim model. Advantages of DeepJudge comparing to traditional watermarking or fingerprinting work include: 1) non-invasive, as it works directly on the model and does not tamper with the training process; 2) efficient, as it only needs a small set of seed test cases and a quick scan of the two models; 3) flexible, i.e., it can easily incorporate new testing metrics or test case generation methods to obtain more confident and robust judgement; and 4) fairly robust to model extraction attacks and adaptive attacks. We verify the effectiveness of DeepJudge under three typical copyright infringement scenarios, including model finetuning, pruning and extraction, via extensive experiments on both image classification and speech recognition datasets with a variety of model architectures.

个人介绍：Jingyi Wang is currently an Assistant Professor in Zhejiang University. Before that, he was a Research Fellow in National University of Singapore. He received his B.E. and Ph.D. from Xi'an Jiaotong University and Singapore University of Technology and Design in 2013 and 2018 respectively. His research concerns how to better design, implement and analyze artificial intelligence (AI) systems and cyber-physical systems (CPS), such as autonomous driving cars, industrial control systems, and recommendation systems, supported by various software engineering (SE) techniques ranging from formal methods, program analysis to software testing. His research has been published in top conferences and journals such as ICSE, S&P, TSE, TACAS and FM. He has won ICSE's Distinguished Paper Award twice (ICSE 2018/2020). His ICSE 2020 AI Fairness Testing work has also been selected for ACM SIGSOFT Research Highlights.

3:30 - 4:10:

Invited Talk 3: **Adaptive Fairness Improvement based on Causality Analysis**, published at ESEC/FSE 2022 *(Mengdi Zhang, Singapore Management University)*

报告摘要：Given a discriminating neural network, the problem of fairness improvement is to systematically reduce discrimination without significantly scarifies its performance (i.e., accuracy). Multiple categories of fairness improving methods have been proposed for neural networks, including pre-processing, in-processing and post-processing. Our empirical study however shows that these methods are not always effective (e.g., they may improve fairness by paying the price of huge accuracy drop) or even not helpful (e.g., they may even worsen both fairness and accuracy). In this work, we propose an approach which adaptively chooses the fairness improving method based on causality analysis. That is, we choose the method based on how the neurons and attributes responsible for unfairness are distributed among the input attributes and the hidden neurons. Our experimental evaluation shows that our approach is effective (i.e., always identify the best fairness improving method) and efficient (i.e., with an average time overhead of 5 minutes).

个人介绍：Mengdi Zhang is a third year Ph.D candidate in Singapore Management University supervised under Prof Jun Sun. She holds a B.E in the University of Electronic Science and Technology of China. Her research interests are mainly on AI security, including machine learning interpretability and fairness testing.

4:10 - 4:50:

Invited Talk 4: **Robustness Analysis for DNNs from the Perspective of Model Learning**, published at ICSE 2022 *(Pengfei Yang, Chinese Academy of Sciences)*

报告摘要：To analyze local robustness properties of deep neural networks (DNNs), we present a practical framework from a model learning perspective. Based on black-box model learning with scenario optimisation, we abstract the local behavior of a DNN via an affine model with the probably approximately correct (PAC) guarantee. From the learned model, we can infer the corresponding PAC-model robustness property. The innovation of our work is the integration of model learning into PAC robustness analysis: that is, we construct a PAC guarantee on the model level instead of sample distribution, which induces a more faithful and accurate robustness evaluation. This is in contrast to existing statistical methods without model learning. In the experimental evaluation, our method outperforms the state-of-the-art statistical method PROVERO, and it achieves more practical robustness analysis than the formal verification tool ERAN.

个人介绍：Pengfei Yang is working as a post-doc in Institute of Software, Chinese Academy of Sciences. He mainly works on AI safety and probabilistic model checking. In the domain of AI safety, he proposes varieties of methods including symbolic propagation in abstract interpretation, verification through Lipschitz constants, spurious regions guided refinement, and PAC-model learning based verification technique, and he also participates in developing the first Chinese platform for DNN verification --- PRODeep. Besides these, Pengfei Yang is also interested in probabilistic model checking, probabilistic programs, quantum computation and reinforcement learning.

4:50 - 5:00:

Closing Remarks *(Sun Jun, Singapore Management University)*