YUHAO MAO

D-INFK, ETH Zurich, Switzerland
+41 764428769 | e: myh821746176@outlook.com

EDUCATION

ETH Zurich Zurich, Switzerland
Master in Computer Science September 2021 – Present

Zhejiang University Hangzhou, China
Bachelor in Applied Mathematics & Finance September 2017 – July 2021

- GPA: 3.94/4.00
- Selected Courses: Mathematical Analysis II (98), Ordinary Differential Equations (98), Stochastic Processes (96), Multivariate Statistical Analysis (97), Data Structure & Algorithm Analysis (Fundamental Course: 95, Advanced Course: 91).
- Admitted to Chu Kochen Honors College (top 5% freshmen admitted annually)

Massachusetts Institute of Technology Boston, USA
Machine Learning Summer Program July 2019

- Participate in immersive academic courses in traditional machine learning, deep learning, and reinforcement learning. Lead a diverse and multi-institution team of students to complete a project on artist style classification.
- Utilizing ensemble learning and data augmentation, we developed a model that achieved almost the same performance as the state-of-art but with much simpler architecture. We were awarded best team performance and a final score of 96/100.

JOB EXPERIENCE

Higgs Asset Hangzhou, China
Internship in Quant Research March 2021-July 2021
Topic: Rank Prediction of the Future Return

- Found multiple patterns which systematically improves the prediction model.

RESEARCH EXPERIENCE

Zhejiang University, College of Computer Science and Technology Hangzhou, China
Research Assistant to Professor Shouling Ji, ZJU 100-Young Professor December 2020 – Present
Topic: Statistical Test for Neural Explanations

- While neural explanations are designed to explain neural networks, no certification of their stability is provided. In this work, we propose a toolbox of statistical tests to certify the stability of neural explanations. We show an application of this method greatly improves the stability without loss of fidelity and increases the robustness against adversarial attack.
- Working paper as a co-first author.

The Helmholtz Center for Information Security Saarland, Germany
Research Assistant to Dr. Yang Zhang August 2020 – June 2021
Topic: Exploiting denoising malfunction of neural networks to inject a backdoor without poisoning

- Most successful existing backdoor attacks to neural networks involve poisoning, which run a multi-object optimization to make models deceive users when a trigger is present. We illustrate that without poisoning, the feature extractors of neural networks already preserve enough irrelevant information which can be further exploited to inject a neuron-level backdoor.
- Independent research, under the supervision of Zhang, to develop a neuron-level ensemble-based classifier with careful mathematical design. 100% separability between trigger-present and trigger-absent images achieved with only ten neurons and a trigger consisting of eleven pixels in the dataset CIFAR10. This work is presented as my Bachelor's degree thesis.

The Helmholtz Center for Information Security Saarland, Germany
Research Assistant to Dr. Yang Zhang July 2020 – August 2020
Topic: Efficient and Generalized Artificial Brain Stimulation for Detecting Backdoors in Neural Networks

- Recent studies have shown that a neural network can be trojaned to inject backdoors, *i.e.*, a normal behavior on benign inputs but making malicious decisions when a trigger is invoked. This project improved a recently proposed novel defense enabling it to work efficiently on large networks (scanning and analysis currently takes hours). This work generalized techniques that, until now, only studied a limited range of backdoors, and has been used to study standardly trained models to discuss natural backdoors.

- Independent research, under the supervision of Zhang, to accelerate the technique on large networks by quadratic magnitudes using layer-subsampling (*e.g*. approximately 100x for ResNet-50 without loss of detection precision).

Zhejiang University, College of Computer Science and Technology Hangzhou, China
Research Assistant to Professor Shouling Ji, ZJU 100-Young Professor December 2019 – June 2020
Title: Transfer Attacks Revisited: A Large-Scale Empirical Study in Real Settings

- Neural networks are vulnerable to crafted inputs, known as adversarial examples (AEs), that possess a mysterious property called transferability, *i.e*., AEs crafted to fool one network are likely to also fool another independent model. This project studies that property in real-world settings, contrasting and extending previous studies set in simplistic and unrealistic lab settings.
- We empirically consider the following questions: 1) Are real systems vulnerable to transfer attacks? 2) Which attack transfers better in real settings? 3) How is the transferability influenced by surrogate settings? 4) How do sample-level properties contribute to the transferability? To emphasize, we overturn two conclusions previously made in lab settings and extend many others.
- First author to a paper summited, conducting the majority of experiments, analyses, visualizations and paper preparation.

ADDITIONAL INFORMATION

Additional Professional and Extracurricular Experiences

- Deliver oral presentation and attend lectures as one of four specially invited students at the *4th Annual Honors International Faculty Institute Workshop* at Texas Christian University, USA, in June 2019.
- Spend approximately 200 hours volunteering as an undergraduate student, including as an assistant for the *11th International Chinese Statistics Association (ICSA) International Conference*.

Interests

- Teaching math classes at elementary schools, high schools and as a part-time calculus tutor to first-year university students.
- During the first half of 2020, achieved a 20% yield rate in the funds that I invested in.
- Member of the CKC College volleyball team for the past three years.

Computer and Language Skills

- Fluent: Python (95/100 course score, multiple course and research projects in Python, familiar with Pytorch, Pandas, Matplotlib, etc.), Latex, C, R, Markdown
- Experienced: MATLAB, HTML, SQL