Yichi Zhang

✉ yichics02@gmail.com | ☎ +86-138-8430-8849 | ⌂ yichics.github.io | GitHub

Education

Zhejiang University Sept. 2021 – Jun. 2025

- B.Eng. in Information Engineering, GPA: 3.92/4.0
- Minor in Computer Science and Technology
- **Honor**: Third-class Scholarship, Zhejiang University

Publications

Diffusion Noise Feature: Accurate and Fast Generated Image Detection
Yichi Zhang, Xiaogang Xu
Submitted to International Journal of Computer Vision (IJCV), [Paper]/[Code]

Revisiting Defense Mechanisms in FL: Effective and Efficient Backdoor Attack via Trigger Pre-optimization
Xi Chen, Rui Zeng, Yichi Zhang, Chunyi Zhou, Yuwen Pu, Qingming Li, Lu Zhou, Zhe Liu, Shouling Ji
Submitted to IEEE Symposium on Security and Privacy 2025 (IEEE S&P 2025), [Paper]/[Code]

Research Experience

AI-Generated Content Detection Jul. 2023 – Mar. 2024

Advisor: Prof. Xiaogang Xu (Zhejiang University & CUHK)

- Proposed Diffusion Noise Feature, a novel representation that leverages estimated noise from inverse diffusion process to construct image features, pioneering its application in generated image detection.
- Developed a comprehensive evaluation framework for generated image detection, covering detection accuracy, generalization of unknown model-generated images, and robustness to perturbations.
- Achieved state-of-the-art detection performance compared with other detection methods, especially addressing the challenges of detecting images generated by diffusion models.

Backdoor Attack in Horizontal Federated Learning Jul. 2023 – Mar. 2024

Advisor: Prof. Shouling Ji (Zhejiang University)

- Proposed PreFed, which is the first to investigate the relationship between client dataset distribution and model updates in federated learning, enabling the design of novel and efficient backdoor triggers.
- Derived a theoretical bound to express the relationship between the attacker's backdoor attack boundary and the IID degree of the dataset label distribution in federated learning.
- Achieved state-of-the-art attack success rates while bypassing various defense mechanisms.

Jailbreaking Attacks in Large Language Model Inference Jul. 2024 – Aug. 2024

Advisor: Prof. Ting Wang (Stony Brook University)

- Re-examined jailbreaking attacks, focusing on uncovering security vulnerabilities in LLM safety alignment mechanism when faced with near-natural language jailbreak prompts.
- Developed a LLM jailbreaking attack by using controllable decoding to generate adversarial suffixes, fine-tuning a jailbreak prompt generator to achieve efficient and low-perplexity LLM jailbreaking.

Corpus Poisoning Attacks in Retrieval-Augmented Generation System Aug. 2024 – Oct. 2024

Advisor: Prof. Ting Wang (Stony Brook University)

- Evaluated the robustness of various RAG systems against corpus poisoning attacks, finding that Graph RAG, due to its graph indexing mechanism, exhibits strong resistance to poisoning.
- Developed a poisoning attack targeting Graph RAG by modifying the graph index through prompt injection, achieving a higher attack success rate with minimal malicious text.

Dynamic 3D Point Cloud Compression Nov. 2024 – Now

Advisor: Prof. Qianqian Yang (Zhejiang University)

- Proposed Implicit Neural Representations for point cloud compression for the first time, instead of traditional encoder-decoder structures or conventional compression methods.
- Employed KAN to separately learn the existence and attributes of point clouds, combining sparsification learning and pruning techniques to efficiently compress complex point clouds with fewer parameters.
- Achieved state-of-the-art signal-to-noise ratio and compression ratio across multiple benchmarks.

Project

PEFT in Few-Shot Learning Mar. 2024 – Jun. 2024

Advisor: Prof. Xiaojin Gong (Zhejiang University)

- Reviewed and summarized over 20 state-of-the-art papers on PEFT or Few-Shot Learning.
- Reproduced various PEFT methods to validate their effectiveness in classification tasks.
- Developed a novel PEFT method combining LoRA and Adapter for few-shot learning.

Backdoor Attacks and Defenses in Machine Learning Oct. 2023 – Jan. 2024

Advisor: Prof. Jake Zhao (Zhejiang University)

- Reviewed and summarized over 40 state-of-the-art papers on backdoor attacks, defenses.
- Reproduced backdoor attack and defense experiments using a custom-designed backdoor toolbox.
- Conducted a comprehensive survey of 30+ state-of-the-art papers on LLM safety and agent safety.

Parallel Solving for PDN Analysis Using Machine Learning Mar. 2023 – Jun. 2023

Advisor: Prof. Cheng Zhuo (Zhejiang University)

- Use sparse matrix partial reordering and Cholesky decomposition to optimized the PDN solver.
- Parallelize subgraph block calculations during the Schur complement computation.
- Optimize resource allocation using machine learning to improve approximation accuracy.

Skill

Languages: English (Advanced, TOEFL: 105, Reading 30, Listening 26, Speaking 23, Writing 26), Mandarin (Native)

Programming: Python, C, C++, Verilog, Shell, HTML, MATLAB, LaTex, SPICE, SQL

Software Tools: Linux, MacOS, Windows, MS Office Software, MATLAB, Altium Designer, ModelSim, Vivado, Advanced Design System, CST Studio Suite, GNURadio, Multisim, HSPICE, SQL Server Management Studio

Soft Skills: Communication skills, Presentation skills, Sense of responsibility, Organizational abilities

Research Interests

My research interests primarily focus on **trustworthy machine learning**. My long-term goal is to advance the development of secure and responsible machine learning systems for real-world applications. Additionally, I am interested in enhancing the transparency, fairness, and generalization of machine learning, which involves both the theory and foundational models of machine learning. At the same time, I explore the intersection of machine learning with other fields, such as cybersecurity and EDA.