Zihan Jiang

Zhejiang, China | Email: zihan235711@gmail.com | Tel: + (86) 18856027279

Education

Zhejiang University 09/2022-06/2026

- **B.S in Electronic Science and Technology | GPA: 3.56/4.0** *Zhejiang,China*
- **Core Courses:** *Digital System, Introduction to Information & Electronic Engineering, IOT System Design, Edge computing development practices, Physics Fundamentals of Information Electronics*
- **Honors:** *Academic Scholarship in Year 2023; 1st Prize in National Undergraduate Electronics Design Contest*

Internship Experience

Exploiting Representation Engineering-based Defense against Inference Cost Attacks 07/2025–11/2025

Summer Intern, UCLA — Los Angeles, CA, USA

- Reproduced multiple inference-cost attack techniques inlcuding Verbose, Hidden-tail, VLMInferSlow and Engorgio to force multi-model pipelines into excessive generation and saturate server GPU usage.
- Investigated multi-model vulnerability to Inference Cost Attacks in white-box settings by combining VLM resource-consumption attacks with LLM DoS methods.
- Designed a representation engineering-based defense framework that redirects attacked input distributions toward benign ones and generalizes across attack patterns.
- Apply the attack-defense pattern to different multi-model code basis such as Qwen and BLIP, proving the fact that current models share the same vulnerability to such attacks and can be enhanced with our defense frame.
- Achieved 95.3% benign outputs and reduced 4.3% of attacked samples by over half in generation length on Qwen, where Hidden-tail originally induced >57% of samples to hit maximum length.

Academic Research

Enhancing Image Classification Network Robustness Against Adversarial Attacks with MAML 11/2024–06/2025

- Developed a novel attack-aware defense approach: proactively embedded "controlled vulnerabilities" into the model to mislead adversarial attackers towards predictable patterns, enabling targeted defense.
- Implemented gradient-based adversarial attack generation with **i-FGSM, DeepFool** and **C&W**, observed >95% attack success rate, validating the vulnerability exploitation phase.
- Designed an ISP-based pre-network module optimized with **Adam**(Adaptive Moment Estimation), neutralizing adversarial samples aligned with induced vulnerabilities, yielding 85% accuracy and 84.3% ASR reduction on adversarial inputs.
- Integrated **MAML**(Model-Agnostic Meta-Learning) for meta-training initial weights across multi-attack tasks, enhancing robustness generalization and forcing attackers into specific predictable patterns.

Reproduction of VCD & Activation Steering Decoding (CVPR 2024 / ACL 2024) 11/2024–06/2025

- Reproduced hallucination-mitigation methods from VCD: Mitigating Object Hallucinations in Large Vision–Language Models through Visual Contrastive Decoding *CVPR 2024, Poster Highlight* and Activation Steering Decoding: Mitigating Hallucination in Large Vision–Language Models through Bidirectional Hidden State Intervention ACL 2024 using PyTorch for hidden-state extraction and visualization.
- Implemented an **end-to-end correction pipeline** computing hidden-state contrast vectors between benign and adversarial samples, and applied them via model hooks to steer model outputs toward reliable, low-hallucination responses.
- Resolved model compatibility and efficiency issues by developing a **unified hook-based framework**, enhancing reproducibility and robustness across different VLM architectures.

Exploring Backdoor Attacks in Embedded Sensor Systems 02/2024-05/2024

- Investigated the feasibility of hardware backdoor attacks targeting sensor components in embedded systems, drawing inspiration from software fuzzing techniques and hardware-based exploits such as DoS attacks and Rowhammer.
- Proposed an attack framework for injecting malicious parameters into a sensor's internal pipeline, manipulating color correction matrices and edge enhancement algorithms to induce perception errors in camera and microphone.
- Explored the potential for leveraging electromagnetic emissions to trigger hidden backdoor mechanisms, demonstrating the risks of unverified sensor security in Cyber-Physical Systems, including attacks on control, IoT, and recognition CPS.

Course Projects

Edge Computing: Music Genre Classification and Personalized Song Generation Tools 9/2024-10/2024

- Trained a **Deep Neural Network** to classify 10 music genres using a dataset of 200 30-second audio samples.
- Deployed the model on an edge device **Arduino Nano 33 BLE** under strict memory constraints for real-time genre recognition.
- Integrated cloud functionality via **Suno AI API** to generate personalized songs based on classified genres.
- Achieved end-to-end song generation with **10s** latency after input classification.

FPGA-Based Smart Pillbox for Elderly Medication Management 04/2024-06/2024

- Developed an FPGA-based smart pillbox featuring precise timing control, real-time tracking, and cloud synchronization with doctors' prescriptions.
- Integrated HMI, sensor-based medication verification, and multi-device connectivity to support caregiver monitoring and elderly safety.

Skills

- **Programming Languages:** C/C++, Python, Java, Verilog, Assembly, VHDL, Matlab
- **Circuit Design Software:**Altium Design, AutoCAD, JLCPCB, CST
- **Simulation&Testing:**Modelsim, LTspice, TINA, Multisim, STM32CubeMX, Keil, Vivado