Electrical & Computer Engineering

S E M I N A R

Louisiana State University

Towards Robust and Secure Deep Learning: From Algorithmic Hardening to Hardware-Aware Defense

Ruyi Ding

Northeastern University

Abstract—As artificial intelligence systems become increasingly pervasive, securing them demands a holistic approach that spans learning algorithm to deployment hardware. In this talk, I will present a layered defense framework that encompasses optimization algorithm design, model architecture pruning, and protection mechanisms leveraging hardware-level signals. We first tackle applicability authorization-protecting pre-trained model's IP from unauthorized transfer-by designing EncoderLock, a systematically method that blocks malicious probing. By embedding task-specific authorization into pre-trained encoders, it ensures models restrict illegitimate classification heads while maintaining intact performance on benign ones. However, optimization-centric protections alone are insufficient for locally deployed models. To fortify security at the model architectural level, we introduce Non-Transferable Pruning, which transforms efficiency-driven pruning into a defense mechanism, hardening model's IP.

Yet, even robust algorithmic defenses remain vulnerable to high-priority adaptive attacks. To further enhance model robustness, our design incorporates hardware-level defenses. EMShepherd exemplifies hardware-software co-design: by analyzing electromagnetic emissions from DNN accelerators, it detects adversarial inputs in real time-without relying on model internals. This hardware-informed approach complements algorithmic safeguards, creating a unified defense where physical-layer observability reinforces software resilience.

By combining these complementary strategies, my work demonstrates that robust AI security relies on a harmonious, multi-layered defense. I conclude by envisioning future AI systems that combine hardware-based defenses with robust algorithms to secure deployment against evolving adversarial threats.

When: Tuesday, **11 February 2025,** 9:30 - 10:30
Where: Room 3316E Patrick F. Taylor Hall
Info: `https://www.lsu.edu/eng/ece/seminar`
Food: ***Coffee and cookies will be served.***