COMPUTER SCIENCE & ENGINEERING
UNIVERSITY OF MICHIGAN

DISSERTATION DEFENSE

Neal Mangaokar

Evaluating and Designing Defenses against Input Perturbation Attacks on Black-Box Machine-Learning Models

Tuesday, May 20, 2025
2:00pm – 4:00pm
2311 EECS / Hybrid – Zoom

ABSTRACT: Machine learning (ML) models are vulnerable to adversarial input perturbations, which can lead to misclassifications or harmful outputs. Unfortunately, the extensive and early defense efforts from the community --- such as obfuscated gradients, adversarial training, and black-box restrictions --- have failed against adaptive adversaries. Nonetheless, vulnerable ML models are still being deployed in critical applications, such as deepfake detection and LLM-powered chat clients.

Recently, a new class of defenses introduces external guard components to detect and reject adversarial inputs, including stateful defenses for classification tasks and Guard Models for Large Language Models (LLMs). These modern defenses claim state-of-the-art and significantly reduced attack success rates, but their robustness against adaptive attacks remains uncertain. This thesis thus works towards the following two questions: (a) do these modern guard-based defenses also fail against an adaptive attacker? (b) If so, what recourse do developers have to defend their models? This thesis critically evaluates these defenses and explores alternative approaches to improving adversarial robustness.

First, we assess stateful defenses for ML-as-a-Service (MLaaS) applications and introduce OARS, an adaptive black-box attack that effectively bypasses these defenses, increasing attack success rates from nearly 0% to almost 100% across multiple datasets. Next, we evaluate Guard Models in LLMs and propose PRP, a novel prefix-based attack that circumvents output moderation, demonstrating vulnerabilities in both open- and closed-source LLM guard implementations.

We also investigate designing more robust defense alternatives. Since general-purpose adversarial defenses have repeatedly failed, we investigate whether robustness can be improved in specific application domains. We focus on deepfake detection, where adversarial perturbations can render detectors ineffective. To address this, we introduce D4 (Disjoint Diffusion Deepfake Detection), an ensemble-based approach that leverages frequency domain redundancy to enhance robustness against black-box adversarial perturbations. D4 significantly outperforms existing defenses in detecting adversarially perturbed deepfakes across various generative techniques.
These findings highlight critical weaknesses in modern guard-based defenses and suggest the importance of domain-specific strategies for improving adversarial robustness.

CHAIR: Prof. Atul Prakash