COMPUTER SCIENCE & ENGINEERING
UNIVERSITY OF MICHIGAN

DISSERTATION DEFENSE

Won Park

Crafting Machine Learning Defenses against Adversaries

Monday, March 27, 2023
12:00 – 1:30pm
2725 Beyster
Hybrid – Zoom

ABSTRACT: Machine learning systems are becoming widely adopted and ubiquitous. Not only are there a growth of products in which machine learning is at their core like autonomous vehicles, but even traditional companies in fields such as finance, telecommunications, and travel are integrating machine learning into their internal structure.
However, like any system, machine learning platforms are prone to security risks and vulnerabilities. Coupled with an ever-accelerating deployment and usage of machine learning systems, the attackers' chance of success and capability of damage increases just as rapidly. What is especially concerning is the large surface area of the machine learning pipeline that is available for attack - from training all the way to inference. With such a wide variety of attack combinations possible, there remains a need to address and explore the many types of attacks and defense that are possible in a machine learning environment.
To address this goal, in this dissertation, we explore some of the different types of security vulnerabilities and attacks that are possible with different types of machine learning systems.

At the inference level, my dissertation explores the possibility of crafting adversarial examples on multimodal sensor fusion models. We also explore a new technique that can be used for defending against adversarial examples: adversarial fine-tuning. On the training side, we explore a gap in the study of attacks during the training phase of the model (backdoor attacks), by exploring the frequency domain of images and how that could affect attacks and detection defenses. Finally, through a collaboration at Ericsson, Inc., we explore how a machine learning framework can be deployed to detect anomalous data while still being cognizant of industry restrictions and metrics.

CHAIR: Prof. Z. Morley Mao