CS 5914 Defending Against ML-powered Adversaries

Fall 2022

1 Course description

The cybersecurity threat landscape is changing as we make significant advances in machine learning (ML) and AI technologies. Can an ML-powered adversary making existing threats harder to defend and introduce new threats? In this work, we will focus on algorithmically intelligent adversaries who can harness AI/ML technologies to launch hard-to-defend attacks against online platforms and their users. Example security and privacy threats include: (1) Attacks that leverage deep generative models to create convincing synthetic "fake content" (e.g., videos, images, text) to mislead users. (2) ML technologies powering large-scale online surveillance, e.g., facial recognition. (3) ML-based attacks that inject hard-to-detect backdoors in ML systems today. (4) ML-based attacks that craft adversarial inputs to violate the integrity of ML systems. (5) ML-based attacks that violate the confidentiality of ML systems. This course is designed for students with a background in machine learning, and who are interested in learning about the implications of ML advances in the security/privacy space.

2 Reference materials

Most reading material will be drawn from research papers published at venues such as IEEE S&P, Usenix Security, CCS, NDSS, IMC, WWW, ICML, and NeurIPS.

3 Prerequisites

Students are expected to have a basic understanding of deep learning and machine learning in general. Familiarity with frameworks like TensorFlow and PyTorch is required. Students who enroll for the course are expected to be highly motivated to learn and work hard and be ready to make up for any prerequisite deficiencies they may have.

4 Grading

This course will be project-based. Grading will be based on class participation, project presentations, project report, and other project deliverables.