CS 6604 Foundation Models and Security

Spring 2024

1 Course description

Foundation models are deep neural networks trained on broad data (e.g., text, images, audio) which can be further adapted for a variety of downstream tasks. Popular examples include Large Language Models (LLMs) and vision models like CLIP and ViT. These models are transforming the way we apply machine learning in various domains. A single foundation model (e.g., LLM) can be used for a variety of downstream tasks, even without any task-specific training. As foundation models become more widely available, it is important to understand how this impacts the cybersecurity landscape. This course will cover advanced topics at the intersection of deep learning and security. This course is designed for students who are interested in learning data-driven security topics that are primarily based on methods from machine learning. We will investigate the implications of foundation models in the security domain.

- Understanding failures of foundation models that can cause harm to its users. This includes:
  - Toxicity in LLM-based applications, e.g., chatbots.
  - Hallucination by LLMs and their inability to provide factual, accurate information.
  - Challenges with filtering undesired content from Text-to-Image models.
- Attacks that violate the integrity of foundation models to cause harm. We will focus on prompt-injection attacks against LLM-based applications. We will also cover new threats impacting LLM plugins.
- Misuse of foundation models to create harmful content. We will study the threat of synthetic media or "deepfakes" produced using foundation models. We will look at both text and vision modalities and focus on synthetic media detection schemes.
- Applications of foundation models to improve security. We will study how foundation models can boost the performance of different security tasks, include deepfake image detection, and network and application security tasks.

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, AAAI and NeurIPS.

3 Prerequisites

Students are expected to have a strong background in deep learning and machine learning. 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

Final grade will be based on the following components:

- Class participation
- Paper summaries
- Paper presentation
- Research project