Adversarial Machine Learning Syllabus

Instructor: Bo Li (lxbosky@gmail.com)
TA: Boxin Wang (boxin.wang@outlook.com)

1 Objectives

After this course, students will be able to understand security and privacy vulnerabilities of machine learning models, as well as how to make the learning tasks robust from various perspectives.

2 Grading

Criteria	Percent of Grade
Project	60%
(Initial Proposal, Due 9.23)	(5%)
(Status Report, Due 10.28)	(15%)
(Final Report & Presentation, Due 12.14)	(40%)
Paper reading and presentation	30%
(Paper reviews)	(10%)
(Presentation)	(15%)
(Peer rating)	(5%)
Class participation	10%

Note: The presentation is evaluated based on both the content of slides and quality of presentation.

3 Prerequisites

1. All enrolled students must have taken machine learning classes.
2. Projects will require training neural networks with standard automatic differentiation packages (TensorFlow, Pytorch).
3. Tentative Goal: Everyone group in the class should have one top-tier conference paper for your project!

4 Candidate topics for final projects:

1. Attacks against general machine learning models such as 3D reconstruction, BERT, and RL systems.
2. Detection against attacks such as Deepfake.
3. GWAS for AI
4. Theoretically understanding of generative models from the game theoretic perspective
5. Applications of GANs (GAN Zoo)

6. Provable robustness for classifiers against different types of perturbation
7. Differential private graphs, and robust graph neural networks
8. Privacy analysis for generative models
9. Robust reinforcement learning
10. Improve model robustness with unlabeled data via semi-supervised learning
11. Robustness testing for different deep neural networks architectures Robust autoML

5 Reading Materials and Project Topics

Checkout: `https://aisecure.github.io/TEACHING/2019.html`