CS 502 – Directed Studies Course: Adversarial Machine Learning

Total Credits: 3

Instructor: Alex Vakanski, vakanski@uidaho.edu

Semester: Fall 2020 (August 24 – December 18, 2020)

Hybrid course

Course Description
The course introduces students to adversarial attacks and defenses on machine learning models. The particular focus is on adversarial examples in deep learning models, due to its prevalence in modern machine learning applications. Covered topics include evasion attacks against white-box and black-box machine learning models, data poisoning attacks, privacy attacks, defense strategies against common adversarial attacks, generative adversarial networks, and robust machine learning models. The course also provides an overview of explainable machine learning and self-supervised machine learning, with an emphasis on deep learning models.

This course is delivered in a hybrid method. The dates for class meetings are indicated in the Course Outline section. In preparation for the class meetings, the students are expected to read the papers listed as required reading in the Course Outline section.

Textbook
There is no required textbook. The required readings for each week are listed in the Course Outline section.

Learning Outcomes
1. Explain the different types of adversarial attacks against machine learning models.
2. Describe the approaches for improved robustness of machine learning models against adversarial attacks.
3. Implement adversarial attacks and defense methods against adversarial attacks on general-purpose image datasets and medical image datasets.
4. Understand the importance of explainability and self-supervised learning in machine learning.

Prerequisites
Machine Learning or Deep Learning

Grading
Four homework assignments, each worth 25 marks.

Course Outline (Tentative)

Date	Topics, Readings, Assignments
Wednesday August 26	**Lecture 1 (zoom meeting): Introduction to Adversarial Machine Learning**
Wednesday September 2	**Lecture 2 (zoom meeting): Deep Learning Overview**
Wednesday September 9	**Lecture 3: Adversarial Machine Learning in Medical Image Processing** **Required readings:**

	1. Ma et al. (2019) Understanding Adversarial Attacks on Deep Learning Based Medical Image Analysis Systems (pdf) 2. Paschali et al. (2018) Generalizability vs. Robustness: Adversarial Examples for Medical Imaging (pdf) **Optional readings:** 1. Finlayson (2019) Adversarial Attacks against Medical Deep Learning Systems (pdf)
Wednesday September 16	**Lecture 4 (zoom meeting): Mathematics for Machine Learning**
Wednesday September 23	**Lecture 5: Evasion Attacks Against Machine Learning Models** **Required readings:** 1. Carlini et al. (2017) Towards Evaluating the Robustness of Neural Networks (pdf) 2. Xiao et al. (2018) Spatially Transformed Adversarial Examples (pdf) **Optional readings:** 1. Goodfelow et al (2014) Explaining and Harnessing Adversarial Examples (pdf) 2. Szagedy et al. (2014) Intriguing Properties of Neural Networks (pdf) 3. Eykholt et al. (2018) Robust Physical-World Attacks on Deep Learning Models (pdf) **Due:** Assignment 1
Wednesday September 30	**Lecture 6 (zoom meeting): Evasion Attacks Against Blackbox Models** **Required readings:** 1. Brendel et al. (2017) Decision-Based Adversarial Attacks: Reliable Attacks Against Black-Box Machine Learning Models (pdf) 2. Bhagoji et al. (2017) Exploring the Space of Black-box Attacks on Deep Neural Networks (pdf) **Optional readings:** 1. Papernot et al. (2016) Transferability in Machine Learning: from Phenomena to Black-Box Attacks using Adversarial Samples (pdf)
Wednesday October 7	**Lecture 7: Poisoning Attacks Against Machine Learning Models** **Required readings:** 1. Liu et al. (2018) Trojaning Attack on Neural Networks (pdf) 2. Shafahi et al. (2018) Poison Frogs! Targeted Clean-Label Poisoning Attacks on Neural Networks (pdf) **Optional readings:** 1. Biggio et al. (2012) Poisoning Attacks against Support Vector Machines (pdf) 2. Jagielski et al. (2018) Manipulating Machine Learning: Poisoning Attacks and Countermeasures for Regression Learning (pdf) 3. Mei et al. (2015) Using Machine Teaching to Identify Optimal Training-Set Attacks on Machine Learners (pdf)
Wednesday October 14	**Lecture 8 (zoom meeting – presented by Matt): Defenses Against Poisoning Attacks** **Required readings:**

	1. Wang et al. (2019) Neural Cleanse: Identifying and Mitigating Backdoor Attacks in Neural Networks (pdf) 2. Gu et al. (2019) BadNets: Identifying Vulnerabilities in the Machine Learning Model Supply Chain (pdf) **Optional readings:** 1. Steihhardt et al. (2017) Certified Defenses for Data Poisoning Attacks (pdf) 2. Liu et al. (2016) Robust High-Dimensional Linear Regression (pdf) 3. Munoz-Gonzalez et al. (2017) Towards Poisoning of Deep Learning Algorithms with Back-Gradient Optimization (pdf) **Due:** Assignment 2
Wednesday October 21	**Lecture 9: Privacy Attacks Against Machine Learning Models** **Required readings:** 1. Shokri et al. (2018) Membership Inference Attacks Against Machine Learning Models (pdf) 2. Bhagoji et al (2019) Analyzing Federated Learning through an Adversarial Lens (pdf) **Optional readings:** 1. Hitaj et al. (2017) Deep Models Under the GAN: Information Leakage from Collaborative Deep Learning (pdf)
Wednesday October 28	**Lecture 10 (zoom meeting – presented by Shoukun): Generative Adversarial Networks for AML** **Required readings:** 1. Xiao et al. (2018) Generating Adversarial Examples with Adversarial Networks (pdf) 2. Samangouei et al. (2018) Defense-GAN: Protecting Classifiers Against Adversarial Attacks Using Generative Models (pdf) **Optional readings:** 1. Arora et al. (2017) Generalization and Equilibrium in Generative Adversarial Nets (GANs) (pdf) 2. Arora et al. (2017) Theoretical limitations of Encoder-Decoder GAN architectures (pdf) 3. Yang (2020) Defending against GAN-based Deepfake Attacks via Transformation-aware Adversarial Faces (pdf)
Wednesday November 4	**Lecture 11: Defenses Against Adversarial Attacks** **Required readings:** 1. Xu et al. (2019) Adversarial Attacks and Defenses in Images, Graphs and Text: A Review (pdf) 2. Tramer et al. (2018) Ensemble Adversarial Training: Attacks and Defenses (pdf) **Optional readings:** 1. Madry et al. (2017) Towards Deep Learning Models Resistant to Adversarial Attacks (pdf) 2. Papernot et al. (2016) Distillation as a Defense to Adversarial Perturbations against Deep Neural Networks (pdf) 3. Meng et al. (2017) MagNet: a Two-Pronged Defense against Adversarial Examples (pdf)

Wednesday November 11	**Lecture 12 (zoom meeting – presented by Haotian): Defenses Against Adversarial Attacks – Part II** **Required readings:** 1. Raghunathan et al. (2018) Certified Defenses against Adversarial Examples (pdf) 2. Zhang et al. (2019) Theoretically Principles Trade-off between Robustness and Accuracy (pdf) **Optional readings:** 1. Lamb et al. (2018) Fortified Networks: Improving the Robustness of Deep Networks by Modeling the Manifold of Hidden Representations (pdf) 2. Gowal et al. (2019) On the Effectiveness of Interval Bound Propagation for Training Verifiably Robust Models (pdf) 3. Wong et al. (2018) Provable Defenses against Adversarial Examples via the Convex Outer Adversarial Polytope (pdf) **Due:** Assignment 3
Wednesday November 18	**Lecture 13: Defenses Against Privacy Attacks** **Required readings:** 1. Papernot et al. (2018) Scalable Private Learning with PATE (pdf) 2. Bindschaedler et al. (2016) Plausible Deniability for Privacy-Preserving Data Synthesis (pdf) **Optional readings:** 1. Abadi et al. (2016) Deep Learning with Differential Privacy (pdf) 2. Dwork et al. (2018) Privacy-preserving Prediction (pdf) 3. Nasr et al. (2018) Machine Learning with Membership Privacy using Adversarial Regularization (pdf)
Wednesday December 2	**Lecture 14 (zoom meeting): Explainability in Machine Learning** **Required readings:** 1. Belle et al. (2020) Principles and Practice of Explainable Machine Learning (pdf) 2. Sundararajan et al. (2017) Axiomatic Attribution for Deep Networks (pdf) **Optional readings:** 1. Arrieta et al. (2019) Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI (pdf) 2. Google (2019) AI Explainability Whitepaper (pdf) 3. Montavon et al. (2017) Explaining Nonlinear Classification Decisions with Deep Taylor Decomposition (pdf)
Wednesday December 9	**Lecture 15: Robustness in Machine Learning** **Required readings:** 1. Ilyas et al. (2019) Adversarial Examples Are Not Bugs, They Are Features (pdf) 2. Weng et al. (2018) Evaluating the Robustness of Neural Networks: An Extreme Value Theory Approach (pdf) **Optional readings:** 1. Yang et al. (2020) A Closer Look at Accuracy vs. Robustness (pdf) **Due:** Assignment 4
Wednesday	**Lecture 16 (zoom meeting): Self-supervised Learning**

December 16	**Required readings:** 1. Chen et al. (2020) A Simple Framework for Contrastive Learning of Visual Representations (pdf) 2. Jing et al. (2019) Self-supervised Visual Feature Learning with Deep Neural Networks: A Survey (pdf) **Optional readings:** 1. Oord et al. (2018) Representation Learning with Contrastive Predictive Coding (pdf)