Spring 2023
CS 785 – Towards Trustworthy Machine Learning/Artificial Intelligence

Instructor: Cong Shi
Email: cong.shi@njit.edu
Course website: https://canvas.njit.edu/
Instructor website: https://njit.webex.com/meet/cs638

Office hours: by appointment via email

Description:
The wide deployment of machine learning/artificial intelligence in real-world systems calls for a set of complementary technologies that will ensure that machine learning is trustworthy. This course will systematically introduce new attack surfaces and techniques arising due to the extending pipeline of machine learning. This course will study research studies designing techniques targeting different stages of the machine learning pipeline, including but not limited to training time (e.g., data poisoning attacks, backdoor attacks) and testing time attacks (e.g., adversarial examples). It also includes state-of-the-art mitigation techniques for these attacks.

This course covers the following topics:

- Machine Learning Basics
- Backpropagation
- Training-time security (attack & defenses)
- Testing-time security (attack & defenses)
- Privacy attacks against machine learning

Grading:

Paper Presentation	50%
Course Project	40%
Class Attendance & Participation	20%

Paper Presentation:

- Besides slides of lectures, course material will include research articles from ACM Digital Library (https://dl.acm.org), IEEE Xplore (https://ieeexplore.ieee.org), and arXiv (https://arxiv.org/).
- The project of this course requires the use of Python and open-source machine libraries (TensorFlow: https://www.tensorflow.org/ Links to an external site., PyTorch: https://pytorch.org/).

Paper Presentation:

- Each student is required to present 2 papers.

- The presentation takes the form of oral presentations at a conference.
- 25 minutes (20 minutes for presentation + 5 minutes for Q&A).
- A list of papers will be posted. You may choose one paper from the list. You’re also welcome to choose papers that are not in the list.

Course Project:

- Carried out by a team of 2 students.
- Program to address a security problem of machine learning.
- Choose from any topics related to the security of machine learning.
- Each group is required to give a proposal presentation and a final presentation
- A final report summarizing the problem, related work, methodology, and results need to be submitted.

Course Outcomes:

After completing the course, students will be able to:

- Understand adversarial attacks targeting different phases of machine learning pipeline.
- Describe the optimization procedures and algorithms to realize testing-phased adversarial attacks in vision and audio domains.
- Describe the defenses against adversarial attacks (e.g., filtering, adversarial training).
- Understand more recent training-phase attacks (e.g., data poisoning, backdoor attacks) that inject malicious behaviors into training data.
- Get familiar with recent training-phase attacks in vision and audio domains.
- Describe the threat of training-phase and testing-phase attacks in physical world

Tentative Course Schedule

1/17 (Tuesday)	Course Introduction	
1/19 (Thursday)	Machine Learning Basics	
1/24 (Tuesday)	Backpropagation	
1/26 (Thursday)	Adversarial Examples	
1/31 (Tuesday)	Paper Presentation (2 papers)	
2/2 (Thursday)	Paper Presentation (2 papers)	
2/7 (Tuesday)	Paper Presentation (2 papers)	
2/9 (Thursday)	Paper Presentation (2 papers)	
2/14 (Tuesday)	Backdoor Attacks	
2/16 (Thursday)	Backdoor Attacks	
2/21 (Tuesday)	Paper Presentation (2 papers)	
2/23 (Thursday)	Paper Presentation (2 papers)	
2/28 (Tuesday)	Paper Presentation (2 papers)	
3/2 (Thursday)	Paper Presentation (2 papers)	
3/7 (Tuesday)	Defense against Adversarial Examples	
3/9 (Thursday)	Defense against Adversarial Examples	Project Proposal Due

3/14 (Tuesday)	Spring Recess	
3/16 (Thursday)	Spring Recess	
3/21 (Tuesday)	Paper Presentation (2 papers); Project Proposal (1 project)	
3/23 (Thursday)	Paper Presentation (2 papers); Project Proposal (1 project)	
3/28 (Tuesday)	Paper Presentation (2 papers); Project Proposal (1 project)	
3/30 (Thursday)	Paper Presentation (2 papers); Project Proposal (1 project)	
4/4 (Tuesday)	Adversarial Attacks in Physical World	
4/6 (Thursday)	Adversarial Attacks in Physical World	
4/11 (Tuesday)	Paper Presentation (2 papers); Project Proposal (1 project)	
4/13 (Thursday)	Paper Presentation (2 papers); Project Proposal (1 project)	
4/18 (Tuesday)	Paper Presentation (2 papers)	
4/20 (Thursday)	Final Project Presentation (2 projects)	
4/25 (Tuesday)	Final Project Presentation (2 projects)	
4/27 (Thursday)	Final Project Presentation (2 projects)	
5/2 (Tuesday)	Final Project Presentation (2 projects)	Last Day of Classe

Honor Code:

Academic Integrity is the cornerstone of higher education and is central to the ideals of this course and the university. Cheating is strictly prohibited and devalues the degree that you are working on. As a member of the NJIT community, it is your responsibility to protect your educational investment by knowing and following the academic code of integrity policy that is found at:
http://www5.njit.edu/policies/sites/policies/files/academic-integrity-code.pdf.

Please note that it is my professional obligation and responsibility to report any academic misconduct to the Dean of Students Office. ***Any student found in violation of the code by cheating, plagiarizing or using any online software inappropriately will result in disciplinary action. This may include a failing grade of F, and/or suspension or dismissal from the university.*** *If you have any questions about the code of Academic Integrity, please contact the Dean of Students Office at dos@njit.edu*

Note in particular that cheating on exams, copying homework assignments and exam papers, and plagiarizing (in full or in part) someone else's work is forbidden.

Collaboration of any kind is PROHIBITED in the exams. As part of projects, students must turn in code or work that has fully been written by him/her and no-one else. Any submitted text or code (even few lines) obtained through the Internet or otherwise, or is product of someone else's work, risks severe punishment, as outlined by the University; all parties of such interaction receive automatically 0 and grade is lowered by one or two levels. Likewise for Exams, if applicable. The work you submit must be the result of your own mental effort and you must safeguard it from other parties; if you can't protect your home computer, use a Lab (AFS) machine.