CSE-891: Adversarial Machine Learning (AdvML)

Sijia Liu
Department of Computer Science & Engineering
Michigan State University
East Lansing, MI 48824-1226
liusiji5@msu.edu

Abstract

Welcome to CSE-891-AdvML. The following will summarize the key course facts and course content. Moreover, course prerequisites will be illustrated for ease of course preparation in the review week (01/11/2021-01/15/2021).

1 Course Description

In recent years, adversarial ML is shown to be a key technique that leads to exciting breakthroughs and new challenges of many AI applications. In particular, adversarial robustness (centered at attack and defense) becomes an emerging topic to promote the trust of AI and enables a better understanding of the pros and cons of deep learning systems. More generally, the idea of *learning with adversary* is crucial for expanding the learning capability, ensuring trustworthy decision making, and enhancing generalizability of machine learning and data mining methods. Despite diverse adversarial concepts and applications, they share very similar learning, computation, and optimization foundations. Thus, the *main course goal* is to teach students how to adapt these fundamental techniques into different use cases of adversarial ML in computer vision, signal processing, data mining, and healthcare. Hopefully, the offered course will also help students to motivate research ideas and to produce research publications.

Overview of course content The course will be taught in a lecture-style, and will cover the following topics[1].

- Adversarial robustness evaluation of deep neural networks: why and how.
- Prediction-evasion attacks in machine learning
- Data-poisoning attacks and Trojan deep models
- Adversarial training: From adversarial exploration to defense
- Robustness certification and randomized smoothing
- Adversarial ML by semi- and self-supervision
- Applications of adversarial attacks and defenses in image classification and beyond
- Adversarial ML vs. explainable ML
- Adversarial robustness vs. generalization and fairness

2 Course Facts and Logistics

Basic information

[1] Topics might be subject to minor modifications over the semester

- **Course time**: Tu & Th 10:20 AM-11:40 AM EST (Online course starts from **Tuesday, January 19**)

  Side note: Per President Stanley's Dec. 21 announcement and department announcement, January 11 – 15 is a *review week*. The purpose of this week is to provide students with the course syllabus and course schedule to allow them to plan for the semester, and ask questions about the course.
- **Course location**: Instructor's Zoom
  `https://msu.zoom.us/j/8793925925` (Meeting ID: 879 392 5925)
- **Office hours**: General Q&A on Th 12-1 PM, or by appointment
- **Communication mode**: Email or D2L for assignments and announcements
- **Course lecture format**: Regular lectures and guest lectures on special topics
- **Homework & exam**: Homework assignments (30%), course presentations (30%), final course project and presentation (40%)

3 Course Prerequisites

The course requires general background knowledge on ML/deep learning (DL) such as convolutional neural networks, optimization (e.g., stochastic gradient descent), and signal processing (e.g., estimation/detection theory). The final project might need using Python and PyTorch/TensorFlow for implementing existing baseline methods and possibly new methods invented during the course project. Some useful material related to this course: Foolbox[2] [1] and [2].

4 Online Course Instruction

Zoom:

1. You can use Zoom on a desktop with a camera/mic/speaker, laptop (with same), or even smartphone.
2. To download Zoom on to your device: go to `https://msu.zoom.us` and click the link "Download Zoom" under "Important Links" heading of the left side of the screen.

To Participate in a Meeting:

1. Click the Zoom link you have on your device.
2. Click "Join a Meeting"
3. Enter Meeting ID of host.
4. The host has to be logged on for it to work. If host is not logged in yet, you will get a message. When the host is on, you will be connected.

References

[1] Jonas Rauber, Wieland Brendel, and Matthias Bethge. Foolbox: A python toolbox to benchmark the robustness of machine learning models. *arXiv preprint arXiv:1707.04131*, 2017.

[2] Nicholas Carlini, Anish Athalye, Nicolas Papernot, Wieland Brendel, Jonas Rauber, Dimitris Tsipras, Ian Goodfellow, Aleksander Madry, and Alexey Kurakin. On evaluating adversarial robustness. *arXiv preprint arXiv:1902.06705*, 2019.

[2] `https://github.com/bethgelab/foolbox`