The Andrew and Erna Viterbi Faculty of Electrical Engineering | TECHNION

Correctness Guarantees for Deep Learning (048890)

Winter 2023/2024
Wednesdays 10:30—12:30

Teaching Staff:
Instructor: Dana Drachsler Cohen [ddana@ee.technion.ac.il]
Office Hours: Wednesdays 12:30

Prerequisites: None
Co-requisites / Courses Without Credit: None
Credits: 2 points

Course Goals and Description

Neural networks have shown tremendous success in many domains. At the same time, recent years have shown the simplicity in fooling neural networks by adversarial example attacks. These attacks undermine the reliability of deep learning-based systems. In this seminar, we will learn about these attacks and how to deal with them. In particular, we will learn:

- Methods for proving robustness of neural networks to adversarial attacks.
- Training methods that leverage formal methods with the goal of improving the neural network's robustness.
- Methods for understanding the robustness behavior of neural networks.

We will focus on practical methods, will study tools that implement them, and learn the limitations required to make these methods capable of analyzing deep networks. We will also discuss open research questions. The seminar will cover papers from leading conferences.

Course Topics

List of papers will be published close to the beginning of the semester.

Grading

The grade will be based on:

1. Presentation (60%): An original presentation (written by the student), showing that the student has understood the paper. Part of the grade is based on the presentation quality (slides + oral presentation). Bonus will be given to creative examples or explanations.
2. Quizzes (30%): At the end of every talk, a short quiz will be posted. The grade is the average of all quizzes, except for the two lowest grades.
3. Attendance (10%).

Course Schedule

First two lectures will be given by Dana.

Registration:

If you are interested, please submit the form: https://forms.gle/Lv1vRDqtA6TcJaXj7

For questions, please contact Dana: ddana@ee.technion.ac.il