NJIT

Department of Computer Science

Can We Trust AI? Towards Practical Implementation and Theoretical Analysis in Trustworthy Machine Learning

Kaidi Xu
Northeastern University

Hosted by Vincent Oria

DATE: Monday, February 15, 2021
TIME: 2:30 PM - 3:30 PM
Webex Link: https://njit.webex.com/njit/j.php?MTID=m223723cd85c947cb062a4a9fa2eca63d

https://cs.njit.edu/seminars

Abstract: Deep learning has achieved extraordinary performance in many application domains recently. It has been well accepted that DNNs are vulnerable to adversarial attacks, which raises concerns about DNNs in security-critical applications and may result in disastrous consequences. Adversarial attacks are usually implemented by generating adversarial examples, i.e., adding sophisticated perturbations onto benign examples, such that adversarial examples are classified by the DNN as target (wrong) labels instead of the correct labels of the benign examples. The adversarial machine learning aims to study this phenomenon and leverage it to build robust machine learning systems and explain DNNs.

In this talk, I will present the mechanism of adversarial machine learning in both empirical and theoretical ways. Specifically, a uniform adversarial attack generation framework, structured attack (StrAttack) is introduced, which explores group sparsity in adversarial perturbations by sliding a mask through images aiming for extracting key spatial structures. Second, we discuss the feasibility of adversarial attacks in the physical world and introduce a convincing framework, Expectation over Transformation (EoT). Utilize EoT with Thin Plate Spline (TPS) transformation, we can generate Adversarial T-shirts, a powerful physical adversarial patch for evading person detectors even if it could undergo non-rigid deformation due to a moving person's pose changes. Third, we stand on the defense side and design the first adversarial training method based on Graph Neural Network. Finally, we introduce Linear relaxation-based perturbation analysis (LiRPA) for neural networks, which computes provable linear bounds of output neurons given a certain amount of input perturbation. LiRPA studies the adversarial example in a theoretical way and can guarantee the test accuracy of a model by given perturbation constraints. The generality, flexibility, efficiency and ease-of-use of our proposed framework facilitate the adoption of LiRPA based provable methods for other machine learning problems beyond robustness verification.

Bio: Kaidi Xu is a Ph.D. candidate at Northeastern University. His research mainly focuses on the robustness of machine learning, including adversarial attacks, formal robustness verification and certified defenses. Besides trustworthy machine learning, he also has broad research interests include model compression & acceleration and explainable AI. His research papers are published in various top conferences such as NeurIPS, ICML, ICLR, IJCAI, AAAI, CVPR, ECCV, ICCV, etc. He also received multiple student travel awards at top conferences and serves as a committee member or reviewer in different conferences and journals. Previously

to Northeastern University, He obtained his M.S. and B.S. degrees from the Department of Computer Science at the University of Florida in 2017 and Sichuan University in 2015 respectively.