USC Viterbi
School of Engineering
Ming Hsieh Department of Electrical and Computer Engineering

Seminar

Robust Classification under Sparse Adversarial Attacks

Dr. Payam Delgosha
Research Assistant Professor
Computer Science Department
University of Illinois at Urbana Champaign

Monday, May 8, 2023
10:00am – 11:00am
Virtual Only

https://usc.zoom.us/j/97124212376?pwd=NTd0QzRzSXk3OGlzL0dIdFdXMmZYZz09
Meeting ID: 971 2421 2376 **Passcode:** 750210

Abstract: It is well-known that machine learning models are vulnerable to small but cleverly-designed adversarial perturbations that can cause misclassification. While there has been major progress in designing attacks and defenses for various adversarial settings, many fundamental and theoretical problems are yet to be resolved. In this talk, we consider classification in the presence of L0-bounded adversarial perturbations, a.k.a. sparse attacks. This setting is significantly different from other Lp-adversarial settings, with p >= 1, as the L0-ball is non-convex and highly non-smooth. In this talk, we discuss the fundamental limits of robustness in the presence of sparse attacks. In order to find an upper bound on the robust error, we introduce novel classification methods that are based on truncation. Furthermore, in order to find a lower bound on the robust error, we design a specific adversarial strategy which tries to remove the information about the true label given the adversary's budget. We discuss scenarios where the bounds match asymptotically. Motivated by the theoretical success of the proposed algorithm, we discuss how to incorporate truncation as a new component into a neural network architecture, and verify the robustness of the proposed architecture against sparse attacks through several experiments. Finally, we investigate the generalization properties and sample complexity of adversarial training in this setting.

Bio: Payam Delgosha received his B.Sc. in Electrical Engineering and Pure Mathematics in 2012, and his M.Sc. in Electrical Engineering in 2014, both from Sharif University of Technology, Tehran, Iran. He received his Ph.D. in Electrical Engineering and Computer Sciences from the University of California at Berkeley in 2020. He joined the computer science department at the University of Illinois at Urbana Champaign as a research assistant professor in 2020. He received the 2020 IEEE Jack Keil Wolf ISIT best student paper award.

Host: Dr. Richard M. Leahy, leahy@sipi.usc.edu