Speaker: Jingze Liu (Binghamton University)

Title: Towards Deep Learning Models Resistant to Adversarial Attacks

Abstract: Recent work has demonstrated that deep neural networks are vulnerable to adversarial examples—inputs that are almost indistinguishable from natural data and yet classified incorrectly by the network. In fact, some of the latest findings suggest that the existence of adversarial attacks may be an inherent weakness of deep learning models. To address this problem, we study the adversarial robustness of neural networks through the lens of robust optimization. This approach provides us with a broad and unifying view on much of the prior work on this topic. Its principled nature also enables us to identify methods for both training and attacking neural networks that are reliable and, in a certain sense, universal. In particular, they specify a concrete security guarantee that would protect against any adversary. These methods let us train networks with significantly improved resistance to a wide range of adversarial attacks. They also suggest the notion of security against a first-order adversary as a natural and broad security guarantee. We believe that robustness against such well-defined classes of adversaries is an important stepping stone towards fully resistant deep learning models.