For favour of Posting

THE HONG KONG UNIVERSITY OF SCIENCE & TECHNOLOGY

Department of Mathematics

PHD STUDENT SEMINAR

Towards Adversarial Robustness by Natural Training Using Deep Stable ODE Networks

By

Mr. Yifei HUANG

Abstract

In this seminar I will talk about a provably stable architecture for Neural Ordinary Differential Equations (ODEs) which achieves non-trivial adversarial robustness under white-box adversarial attacks even when the network is trained naturally. For most existing defense methods withstanding strong white-box attacks, to improve robustness of neural networks, they need to be trained adversarially, hence have to strike a trade-off between natural accuracy and adversarial robustness. Inspired by dynamical system theory, we design a stabilized neural ODE network named SONet whose ODE blocks are skew-symmetric and proved to be input-output stable. With natural training, SONet can achieve comparable robustness with the state-of-art adversarial defense methods. In particular, under PGD-20 (ℓ_∞ = 0.031) attack on CIFAR-10 dataset, our method of natural training achieves 89.36% natural accuracy and 61.62% robust accuracy, while a counterpart architecture of ResNet trained with TRADES achieves natural and robust accuracy 85.28% and 23.06% respectively, in the same setting.

Date : 14 May 2020 (Thursday)
Time : 4:00pm – 5:00pm
Zoom Meeting : https://hkust.zoom.us/j/98027512081

All are Welcome!