*For favour of Posting*

**THE HONG KONG UNIVERSITY OF SCIENCE & TECHNOLOGY**

## Department of Mathematics

# PHD STUDENT SEMINAR

## Black-Box Adversarial Attack

**By**


## Mr. Zhichao HUANG



**Abstract**

Current neural network-based classifiers are susceptible to adversarial examples even in the black-box setting, where the attacker could only query the output of the network. We present a new method for black-box adversarial attack. Unlike previous methods that combined transfer-based and scored-based methods by using the gradient or initialization of a surrogate white-box model, this new method tries to learn a low-dimensional embedding using a pretrained model, and then performs efficient search within the embedding space to attack an unknown target network. The method produces adversarial perturbations with high level semantic patterns that are easily transferable. We show that this approach can greatly improve the query efficiency of black-box adversarial attack across different target network architectures. We evaluate our approach on MNIST, ImageNet and Google Cloud Vision API, resulting in a significant reduction on the number of queries. We also attack adversarially defended networks on CIFAR10 and ImageNet, where our method not only reduces the number of queries, but also improves the attack success rate.


**Date : 15 May 2020 (Friday)**
**Time : 3:00pm – 4:00pm**
**Zoom Meeting : https://hkust.zoom.us/j/98874789179**

*All are Welcome!*