**CRUNCH Seminars at Brown, Division of Applied Mathematics**

**Friday – February 16, 2018**

## The Robust Manifold Defense: Adversarial Training using Generative Models

Guofei Pang

We propose a new type of attack for finding adversarial examples for image classifiers. Our method exploits spanners, i.e. deep neural networks whose input space is low-dimensional and whose output range approximates the set of images of interest. Spanners may be generators of GANs or decoders of VAEs. The key idea in our attack is to search over latent code pairs to find ones that generate nearby images with different classifier outputs. We argue that our attack is stronger than searching over perturbations of real images. Moreover, we show that our stronger attack can be used to reduce the accuracy of Defense-GAN to 3\%, resolving an open problem from the well-known paper by Athalye et al. We combine our attack with normal adversarial training to obtain the most robust known MNIST classifier, significantly improving the state of the art against PGD attacks. Our formulation involves solving a min-max problem, where the min player sets the parameters of the classifier and the max player is running our attack, and is thus searching for adversarial examples in the {\em low-dimensional} input space of the spanner.