Thesis: Finite Gaussian Neurons: Defending Against Adversarial Attacks by Making Neural Networks Say “I Don’t Know”

Abstract: Since 2014, artificial neural networks have been known to be vulnerable to adversarial attacks, which can fool the network into producing wrong or nonsensical outputs by making humanly imperceptible alterations to inputs. While defenses against adversarial attacks have been proposed, they usually involve retraining a new neural network from scratch, a costly task.
In this work, I introduce the Finite Gaussian Neuron(FGN), a novel neuron architecture for artificial neural networks.

My works aims to:
• easily convert existing models to Finite Gaussian Neuron architecture,
• while preserving the existing model’s behavior on real data,
• and offering resistance against adversarial attacks.
I show that converted and retrained Finite Gaussian Neural Networks (FGNN) always have lower confidence (i.e., are not overconfident) in their predictions over randomized and Fast Gradient Sign Method adversarial images when compared to classical neural networks, while maintaining high accuracy and confidence over real MNIST images. To further validate the capacity of Finite Gaussian Neurons to protect from adversarial attacks, I compare the behavior of FGNs to that of Bayesian Neural Networks against both randomized and adversarial images, and show how the behavior of the two architectures differs. Finally I show some limitations of the FGN models by testing them on the more complex SPEECHCOMMANDS task, against the stronger Carlini-Wagner and Projected Gradient Descent adversarial attacks.

Committee:

- Professor Michael I. Mandel, Mentor, Brooklyn College
- Professor Rivka Levitan, Brooklyn College
- Professor Ioannis Stamos, Hunter College

Outside member:

- Dr. Andrew Rosenberg, Google