Computer vision algorithms, such as those implementing object detection, are known to be susceptible to adversarial attacks. Small barely perceptible perturbations to the input can cause vision algorithms to incorrectly classify inputs that they would have otherwise classified correctly. A number of approaches have been recently investigated to generate such adversarial examples for deep neural networks. Many of these approaches either require grey-box access to the deep neural net being attacked or rely on adversarial transfer and grey-box access to a surrogate neural network.

In this thesis, we present a new approach to the synthesis of adversarial examples for computer vision algorithms that only requires black-box access to the algorithm being attacked. Our attack approach employs fuzzing with features derived from the layers of a convolutional neural network trained on adversarial examples from an unrelated data set. Based on our experimental results, we believe that our validation approach will enable designers of cyber-physical systems and other high-assurance use-cases of vision algorithms to stress test their implementations.

Major: Computer Science

Educational Career:
Bachelor's of Computer Science, BS, 2014, Florida Atlantic University

Committee in Charge:
Sumit Kumar Jha, Chair, Computer Science
Leavens Gary, Computer Science
Sharma Valliyil Thankachan, Computer Science

Approved for distribution by Sumit Kumar Jha, Committee Chair, on September 20, 2017.

The public is welcome to attend.