SaTC: CORE: Medium: Collaborative:

Using Machine Learning to Build More Resilient and Transparent Computer Systems

Carnegie Mellon University Duke University

Challenges:

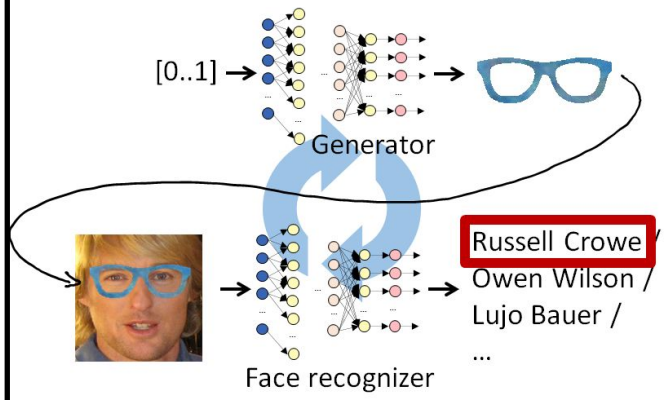
- Develop techniques to better explain the output of ML algorithms (specifically deep neural networks – DNNs)
- Leverage adversarial ML attack algorithms to build systems that are more resistant to attack

Solution:

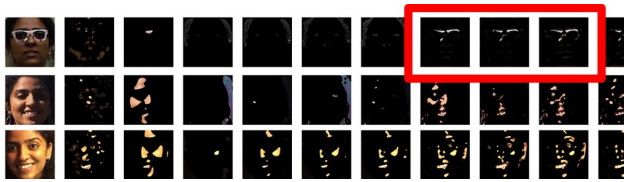
- Model adversaries and defended systems as DNNs within Generative Adversarial Nets (GANs)
- DNNs representing adversary and defended system compete to produce better attacks and a (model of a) more robust system

1801391, Carnegie Mellon University,
Lujo Bauer and Matt Fredrikson
2113345, Duke University, Michael K. Reiter

GANs to create attacks *and* more robust system models



Explanations of DNN outcomes draw attention to adversarial manipulation



Scientific Impact:

- The project will contribute new techniques for explaining DNN outputs and better algorithms for test-time attacks against DNNs
- The project contributes a method for harnessing attacks to design more resilient systems

Broader Impact and Broader Participation:

- More robust DNNs benefit everyone, because DNNs are used in applications from self-driving cars to healthcare to product recommendations
- Visibility of topic and popularity of ML helps attract wide range of students to become interested in research