SaTC: CORE: Medium: Collaborative:

Better defenses through adversarial learning

Challenge:

Use adversarial learning to build more resilient systems

1. Model *attacker* using deep learning (DNNs)
2. Use adversarial ML to find attacker weaknesses
3. Build systems/defenses that take advantage of attacker weaknesses

Technical challenges:

- Practical attacks are subject to constraints typically not supported by current attack approaches
  - e.g., inconspicuousness, well-formedness (of packets, binaries)
- Classifier decisions on adversarial inputs must be explained before they can guide system design

Solution:

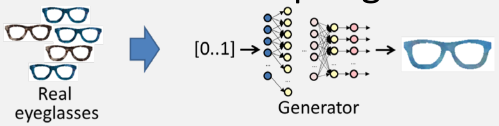
- Develop a general framework for attacks with constraints
  - Thus far: GANs-inspired approach trains a generator to create attack instances that can satisfy multiple constraints, including constraints that cannot be formalized (e.g., inconspicuousness)
  - Ongoing: developing techniques to model constraints for additional domains (e.g., network traffic)
- Use influence-directed explanations to identify input features that caused misclassification
  - Thus far: explanations for attacks on face recognition
  - Ongoing: extending explanation approach to other domains

CMU: award # 1801391, Lujo Bauer, Matt Fredrikson
UNC: award # 1801494, Michael K. Reiter

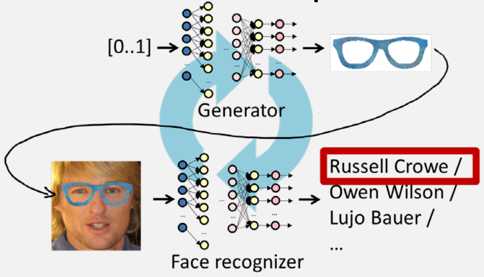
Scientific Impact:

- Better understanding of ML classifier vulnerabilities
- New techniques to explain classifier behavior
- Constructive uses for adversarial inputs and explanations of classifier behavior
- Insights that will lead to more robust classifier designs

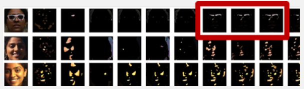
1: train initial input generator



2: augment to produce adversarial inputs



3: explain classifier behavior



Broader impacts:

- DNNs are widely used, including increasingly in applications that impact safety and security (e.g., self-driving aids, critical infrastructure)
- Open source artifacts, practical applications help transition to practice
- Appeal to non-experts entices students to STEM