Adversarial Machine Learning Against Voice Assistant Systems

Supervised by Dr. Yingying (Jennifer) Chen
Team members: David Lau, Celina Zhou, Saurabh Bansal
Meet the Team

Celina Zhou
Duke University
Class of 2022
Major(s): BME, Neuroscience

David Lau
Rutgers University
Class of 2022
Major(s): ECE, CS
Minor(s): Statistics, Economics

Saurabh Bansal
Rutgers University
Class of 2022
Major: ECE
Background

- Voice Assistant Systems
  - User authentication via voice recognition
- Adversarial Attacks
  - Added perturbations to incite misclassifications
Objective

- To study the security of voice assistance systems under adversarial machine learning
- Generate adversarial audio samples to fool voice assistant systems
Methods

- Identify speaker recognition model to attack
  - X-Vector model
    - State-of-the-art speaker recognition model
    - Deep neural network
  - Implemented in TensorFlow, a machine learning framework in Python
Methods

● Untargeted Attack
  ○ Alter audio signal to misclassify as incorrect speaker
  ○ Add a linear perturbation to original signal using Fast Gradient Sign Method (FGSM)

● Targeted Attack
  ○ Change audio signal to imitate a targeted speaker
  ○ If prediction does not match desired speaker, noise is modified to more closely match target speaker
  ○ Targeted attack works iteratively
Results

- Evaluated performance of untargeted adversarial samples on voice assistant system (X-Vector)
Results (cont.)

- Survey to determine the discernable threshold epsilon value

- EPSILON = 1E-05
- EPSILON = 3.16E-05
- EPSILON = 0.0001
- EPSILON = 0.000316
- EPSILON = 0.001
- EPSILON = 0.00316
- EPSILON = 0.01
- EPSILON = 0.0316
- EPSILON = 0.1
Results (cont.)

Targeted attacks:
Future Work

- Effect of room impulse response on attack efficacy
- Disguise attacks
Thank you!

Any questions?