Division Of Work

Group 98

CYBER ATTACKS ON MACHINE LEARNING MODELS: A STUDY OF ADVESARIAL VULNERABILITIES

Group Members

Piyush Gupta	17103067
Chitrank Mishra	17103103
Dharmesh Pratap Singh	17103279

Project Work:

Keypoints of the work done by the team as individual contributions is mentioned as under:

Piyush Gupta

- Collected the requirements for implementing selected two attacks.
- Implemented the Fast Gradient Sign Method and Basic Iterative Method algorithms.

Dharmesh Pratap Singh

- Collected the requirements for implementing selected two attacks.
- Implemented the One Pixel Attack and C&W Attack algorithms.

Chitrank Mishra

- Designed the frontend using Bootstrap and Javascript
- Designed backend server using Django framework.
- Connected all the files and created routes to targeted attack methods.

Research Work:

10 research papers are summarized by each of the group members for the term paper and major project study. They are mentioned as under:

Piyush Gupta

- Poisoning Attacks against Support Vector Machines
- Evasion attacks against machine learning at test time.
- Robust Physical-World Attacks on Deep Learning Visual Classification.
- Practical Black-Box Attacks against Machine Learning.
- Adversarial Examples Are Not Easily Detected: Bypassing Ten Detection Methods
- Adversarial Examples: Attacks and Defenses for Deep Learning
- Adversarial vulnerability for any classifier
- Boosting Adversarial Attacks With Momentum.
- Generating Natural Adversarial Examples
- Learning More Robust Features with Adversarial Training

Dharmesh Pratap Singh

- Fundamental limits on adversarial robustness
- Explaining and harnessing adversarial examples.
- Synthesizing Robust Adversarial Examples
- Adversarial Examples are not Bugs, they are Features.
- Universal adversarial perturbations
- Parseval Networks: Improving Robustness to Adversarial Examples
- Adversarial Examples Are a Natural Consequence of Test Error in Noise
- Are adversarial examples inevitable
- A Simple Explanation for the Existence of Adversarial Examples with Small Hamming Distance
- Evaluating a Simple Retraining Strategy as a Defense Against Adversarial Attacks.

Chitrank Mishra

- Machine Learning in Adversarial Settings
- Adversarial Machine Learning at Scale.
- Accessorize to a Crime: Real and Stealthy Attacks on State-of-the-Art Face Recognition
- Standard detectors aren't (currently) fooled by physical adversarial stop signs
- No Need to Worry about Adversarial Examples in Object Detection in Autonomous Vehicles
- Adversarial Attacks on Neural Network Policies
- Simple Black-box Adversarial Attacks
- APE-GAN: Adversarial Perturbation Elimination with GAN
- A Simple Unified Framework for Detecting Out-of-Distribution Samples and Adversarial Attacks
- Defense against Adversarial Attacks Using High-Level Representation Guided Denoiser.