Adversarial Machine Learning (ML)

Guaranteeing AI Robustness against Deception (GARD)

Project Description

The field of Adversarial Example Research (see resources listed below) provides techniques that can be used to “hack” and deceive machine learning systems in ways that are imperceptible to humans. This will have dire consequences on the use of Machine Learning in applications such as National Security and Healthcare.

In our proposal, we propose using a functional analysis technique called Compressed Sensing as a defense strategy for machine learning systems against Adversarial Examples. We’d like you to learn about the fundamentals of attacks and defenses, with an emphasis on how they’re implemented.

Read through the resources, and be ready to discuss

1. A basic overview of the problem we are studying
2. The code in the Jupyter Notebook github repo provided in the resources section, including the strengths and weaknesses of the authors’ implementation
3. Your own code for implementation of a JPEG Based Preprocessing Defense, and a Backward Pass Differentiable Approximation (BPDA) adaptive attack that breaks the JPEG defense.
    a. Replicate the key result from this notebook. You can implement attacks and defenses yourself with “pure” neural network code in PyTorch, Tensorflow, etc., or you can use any of the frameworks for Adversarial ML that are out there (links in resources section)

Resources

Embedded Intelligence Briefing/Presentation (Task I.4), from January 2021

- **A detailed presentation outlining our research into this area as of Jan 2021**

Athalye, Anish, Nicholas Carlini, and David Wagner. "Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples." International Conference on Machine Learning. PMLR, 2018.

- **https://arxiv.org/abs/1802.00420**

Code supporting Athalye 2018 which shows a JPEG defense and adaptive attack:

- https://github.com/anishathalye/obfuscated-gradients/blob/master/inputtransformations/jpeg.ipynb

Adversarial ML Tutorial

- https://adversarial-ml-tutorial.org/

Adversarial ML Attack Framework Resources

- https://github.com/Trusted-AI/adversarial-robustness-toolbox
- https://github.com/cleverhans-lab/cleverhans
- https://github.com/bethgelab/foolbox