Attack => Adversarial, Data Poisoning

Adversarial

Consist of specially-crafted data points that are routed to the ML model to cause a faulty or wrong inference.

Videos

- Adversarial Attacks on Neural Networks - Bug or Feature?
- Adversarial Examples and Adversarial Training

Papers

- One Pixel Attack for Fooling Deep Neural Attack
- **Adversarial Examples Are Not Bugs, They Are Features** (About Robust Features)

Code

- Robustness Package - Input Manipulation and Robustness of models

Data poisoning

Occur at training time and inject poisoned data points in the dataset.

Possible scenarios

- Scraping images from the web
- Harvesting system inputs (spam detector)
- Federated learning & data collection

Papers

- Feature Collision [Code] (Clean label poisoning attack)
- Backdoor Attack (Fools models by imprinting a small number of training examples with a specific pattern (trigger) and changing their labels to a different target label)
- Bullseye Polytope [Code]
- Targeted Poisoning
  - Label Flipping (Attacker is allowed to change the label of examples)
  - Watermarking (Attacker perturbs the training image, not label, by superimposing a target image onto training images)

Poisoning Benchmark Paper

Code

- **Adversarial Robustness Toolbox** [Both for Attacks and Defenses]

Ensemble Method

Ensemble learning is a machine learning technique that involves combining multiple models to improve the performance of a single model. The idea behind ensemble learning is to leverage the power of multiple models that are individually weak, but together can make better predictions than any single model alone.

Techniques [Bagging, Boosting, Stacking]

Bagging: Training multiple models on different subsets of the data, and then combining their predictions by taking the average or majority vote.

Boosting: Sequentially training models on the same dataset, but each subsequent model focuses on the errors of the previous model.

Stacking: Training multiple models and using their predictions as inputs to a higher-level model.

Papers

Popular Ensemble Methods: An Empirical Study

Ensemble Method Against Data Poisoning

On the Robustness of Ensemble-Based Machine Learning Against Data Poisoning
(Evaluating the robustness of a hash-based Ensemble approach against data**)**

On Collective Robustness of Bagging Against Data Poisoning [Code]