Machine Intelligence Directed Attack Simulator A platform for ML attack and defense experimentation

Background

Machine Learning (ML)

- Constructs a model without being explicitly programmed
- A training phase adjusts model parameters based on data and error
- Inner working of models are extremely complex
- ML models can be exploited

Types of Attacks

- 1. Evasion
 - Degrade a classifier's performance
 - Uses input perturbation or data poisoning
 - Ex: Fooling a network intrusion system

Label: Strawberry



Label: Go-Kart



2. Theft

- Gain illegitimate access to resources
- Ex: Steal a model to run white-box attacks

- Ex: By using model an attacker can steal sensitive training data

/个\

Problem

Researchers need a platform for experimentation with machine learning attack and defense

Solution: MIDAS

A software platform that allows researchers to experiment with a variety of ML attacks

Features

- 1. User-friendly web interface
- 2. Pre-built models
- Ex: Logistic Regression (LR), ResNet CNN
- 3. Multiple datasets
 - Ex: MNIST, EMBER, Imagenet

MIDAS		
Start Jobs Job Evaluation Models Train Model Initiate Attack Attacks Models Datasets		
Search	Selected Attack:	Selected Attack: Iterative Fast Gradient Sign Method Selected Model: ResNet50 CNN Selected Dataset: Gelected Dataset: Attack Options: target_type: target_class: go-kart
Name Description MNIST The MNIST dataset of handwritten 28x28 digits ResNet-compatible 93 images from the ILSVRC2012 validation set that are pre-	Iterative Fast Gradient Sign Method Selected Model: ResNet50 CNN	
ImageNet processed to be compatible with ResNet	Selected Dataset:	
UCI-00422-Wifi Wireless signal strength for room placement determination Pre-processed malware biparies labled malicious or benign	ResNet-compatible ImageNet ir	input_variation:
Displaying (1 - 4) of 4		num_iterations: 30 target_confidence:

initial_class:
strawberry
target alage:
target_class.
go-kart
input_variation:
10
num_iterations:
30
target_confidence:
90
Initiate Attack



CIPHER Lab: Threat Intelligence and Analytics Division



4. Supported Attacks

- LR Input Perturbation
- Text Document Input Perturbation
- LR Model Stealing
- Adversarial Example Generation Attacks
- Ember Data Poisoning



Future Work

- Additional Attack Frameworks
 - Deepfool
 - Foolbox
 - nn_robust_attacks
- User-defined ML models
- User-uploaded data
- Defense methods

Computational Cybersecurity in Compromised Environments

2018 Fall Workshop | September 17-19 | Atlanta, Georgia