KU LEUVEN

CyberSecurity vs AI research: Adversarial AI and using AI for CyberSecurity

CSAI.network CyberSecurity AI F2F public event

Enrique Argones Rúa
email: enrique.argonesrua@esat.kuleuven.be
imec–COSIC, KULeuven

September 14th, 2022

Outline

Introduction

Network intrusion detection

Malware detection

Adversarial AI

Questions and Answers

Outline

Introduction

Network intrusion detection

Malware detection

Adversarial AI

Questions and Answers

Applications of AI in cybersecurity

It may be used wherever a complex task can be automated

- System robustness: Artificial Intelligence for software testing (AIST)
- System resilience: Threat and anomalies detection (TAD)
  - Malware detection
  - Network intrusion detection
- System response

Examples*

- ML trained on user interaction provides a way of understanding local context and knowing what data to focus on
  - ML can be useful for asynchronous user profiling and for measuring deviation from common behaviours as well as going back to much larger data volumes to understand behaviour
- ML can be useful in detecting new anomalies by learning robust models from the data they have been fed with
- ML trained on immutable attacker 'Tactics, Techniques, and Procedures' (TTP) behaviours (those identified in the Mire Attack framework) can support durable and broad attacker detection

*Artificial Intelligence and Cybersecurity, Technology, Governance and Policy Challenges, CEPS Task Force Report, May 2021

Outline

Introduction

Network intrusion detection

Malware detection

Adversarial AI

Questions and Answers

Network intrusion detection

Main considerations

- Complex systems and expert human supervision
- Multiple sources of data
- Not plug and play AI
- Data privacy

Objective

- Economic landscape
  - Limited amount of cybersecurity analysts
  - Human intensive task
- **Objective**: Improve resource usage efficiency by reducing the number of false alarms

Challenges

- Using only data from project partner
- High dimensionality, heterogeneity of data sources, scarce true positives
- Changing environment
- Difficult to obtain and even demonstrate generalization

Research overview

- ML tools trained include random forests and other ensemble methods
- Small false alarms reduction keeping 0 missed threats
- Global *vs/combined with* per-source-technology ML experiments
- Experiments with higher reach can be performed by our industrial partner

Outline

Introduction

Network intrusion detection

Malware detection

Adversarial AI

Questions and Answers

Malware detection

- Multiple tools use ML for MD, some available as AIAAS
- Large MW databases are needed to create such systems (huge knowledge base)
- Public AI-based MW detectors available
- Increased surface attack

Use case: minimizing AI usage through AI

- Economic landscape:
  - AIAAS MWD usage is expensive
- **Objective**: Optimization of AIAAS MWD usage
  - Customers access at least decision-level information from AI
  - Can we detect which samples are easy?
  - Is it feasible regarding latency and computational power?

Approach

- Collection of samples, including general metadata and AI-based decision, used as ground truth
- Train a classifier to emulate AI-based decisions (partial model inversion)
  - Only limited success expected
  - Good enough to characterize the easy/hard examples
- *Easy examples*: we can trust our classifier decision
- *Hard examples*: we cannot trust our classifier decision
  - Hard examples are derived to AI
  - Critical: MW cannot be classified as benign while being considered an easy example
  - Objective: obtain as low benign being classified as hard examples

Results

Some results presented in a Conference paper[†]:

- Usage goals are achieved
- Not clear how well it generalizes for the future, or optimal model update strategy
- Increased surface attack
- Security still relies on the correctness of the AIAAS MWD

[†]Applying machine learning to use security oracles: A case study in virus and malware detection, Davy Preuveneers, Emma Lavens and Wouter Joosen, WTMC 2022, 6 June 2022, Genova - Italy

New developments

- Add more file type specific information
- First analysis done on MS Windows executable files
- On data collection stage

Outline

Introduction

Network intrusion detection

Malware detection

Adversarial AI

Questions and Answers

On Adversarial ML

What is adversarial ML?

- Study of the attacks on machine learning algorithms, and of the defences against such attacks.
  - The adversary has ML capabilities
  - The adversary has certain knowledge of the deployed ML system

On the Adversary

Adversarial model dimensions[‡]

- Goal
- Knowledge of the system
- Capability to modify the underlying data distribution by manipulating individual samples

[‡]Biggio, Battista; Corona, Igino; Nelson, Blaine; Rubinstein, Benjamin I. P.; Maiorca, Davide; Fumera, Giorgio; Giacinto, Giorgio; Roli, Fabio (2014). "Security Evaluation of Support Vector Machines in Adversarial Environments". Support Vector Machines Applications. Springer International Publishing. pp. 105–153.

Examples (1/2)

- Data poisoning
  - Contaminating the training dataset to effectively change the trained model
  - Vector: systems are trained on collected data where the adversary can inject malicious samples
- Byzantine attacks
  - Modify behaviour of some entities involved in collaborative learning to effectively change the trained model
  - Vector: complex ML systems may rely on distributed computation, and some parties may be attacked

Examples (2/2)

- Evasion
  - Modify individual malicious samples to deceive the ML detector
  - Black box (the adversary has limited access to I/O of the ML system) / white box (the adversary knows the ML system)
- Model extraction
  - The adversary probes a black box ML system to extract the *training data* or the model itself (*model stealing*)
  - Partial knowledge of the model can be also advantageous for the adversary

Regarding Evasion Attacks

- Relevant for both use cases (MWD and NID)
  - Two cases:
    - Make malicious samples being classified as benign (security issue)
    - Make benign samples being classified as malicious (resource usage issue)
  - Both cases need to have access, at least partial, to the ML, or to data with a similar training distribution
  - It may be very difficult to handcraft real attacks that can exploit a detected weakness in the classifier input space
  - Difficult to circumvent or prevent

Regarding Poisoning Attacks

- It requires a powerful adversary
- Watch for changes in data distribution

Outline

Introduction

Network intrusion detection

Malware detection

Adversarial AI

Questions and Answers

Questions and answers

Thanks. Questions?