ADVERSARIAL MACHINE LEARNING

Ewa Piatkowska

Center for Digital Safety and Security
AIT Austrian Institute of Technology
ewa.piatkowska@ait.ac.at
MACHINE LEARNING IN SOC

- Security Operations Centers (SOCs) deal with vast amounts of data
- Machine Learning (ML) used to
  - automate and support security analytics
  - support response strategies

- Host Intrusion Detection System (HIDS)
- Network Intrusion Detection Systems (NIDS)
- Security Monitoring (SIEM): security events filtering and alert correlation
- Malware detection, SPAM filtering, Antivirus

- ML will become even more prevalent in security applications due to recent advances in deep learning models
IS MACHINE LEARNING SECURE?

- Machine learning algorithms can be a target of an attack
- Accuracy ≠ Reliability & Security
- Barreno et al. (2006) Can Machine Learning Be Secure?
  - Can an adversary manipulate a learning system to permit a specific attack?
  - Can and adversary degrade the performance of the learning system to the extent that it is no longer trusted?
  - What techniques can be used to confuse a learning system?
THREAT MODEL & TAXONOMY
THREAT MODEL & TAXONOMY

- **Attack Surface**
  - Acquisition
  - Storage
  - Communication
  - Processing
    - ML Model

- **Stage**
  - Training
  - Inference

- **Knowledge**
  - White box
  - Black box

- **Target**
  - Integrity
  - Availability

- **Influence**
  - Causative
  - Exploratory

- **Specificity**
  - Targeted
  - Indiscriminate
ADVERSARIAL EXAMPLES

“adversarial examples are inputs to machine learning models that an attacker has intentionally designed to cause the model to make a mistake”


Original image + Adversarial Noise = Adversarial Example

PANDA
57.7% confidence

GIBBON
99.3% confidence

Imperceptible to human observers → stealthy attack
CRAFTING ADVERSARIAL EXAMPLES

• A box-constraint optimization problem

\[
\min_{x'} J(f(x'), l') \\
\text{s.t. } ||\eta|| \leq \epsilon, \ f(x) = l, l \neq l'
\]

- Optimization objective function is the distance of targeted prediction \((l')\) score from the original prediction \((l)\) score.
- Constraint on adversarial perturbation \(||\eta|| \leq \epsilon\)

• **Why do adversarial examples exist?**
  - Linear behaviour in high-dimensional spaces

\(f(\cdot)\) model
\(J(\cdot)\) loss function
\(x\) original input
\(x'\) adversarial example
\(l\) correct label
\(l'\) adversarial target label
\(\eta\) adversarial perturbation
TRANSFERABILITY AND UNIVERSAL PERTURBATION

• Adversarial examples can be quickly found by changing the relevant features
  • features favoured by classifiers to discriminate between classes


• Different classifiers tend to rely on the same relevant features
  • adversarial examples generalize well to other architectures and data
  • enable black-box attacks

• Universal Adversarial Perturbations (UAP)
  • finding perturbations which can be applied to different inputs from the dataset and cause the misclassification in the target model

SECURITY AND SAFETY CONCERNS

AUTONOMOUS VEHICLES

• Safety of self-driving cars
  • Pedestrian detection
  • Road sign recognition

Original

Adversarial Example

Original

Adversarial Example


PHYSICAL SECURITY

• Face recognition
  • Adversarial glasses

Sharif et al. (2016), Accessorize to a Crime: Real and Stealthy Attacks on State-of-the-Art Face Recognition


• Generative Adversarial Networks (GANs) used to produce Master Prints, i.e. samples which match many real fingerprints
VIRTUAL ASSISTANTS

- Automatic Speech Recognition for Virtual Assistants
  - Amazon Alexa, Apple Siri, Microsoft Cortana, Google Assistant

MALWARE DETECTION

• MalGAN using GANs to generate malware samples

Hu and Tan (2017), Generating Adversarial Malware Examples for Black-Box Attacks Based on GAN (MalGAN), https://arxiv.org/abs/1702.05983

• Evading MalConv (CNN) by adding few padding bytes

Kolosniaji, Biggio, Roli et al., Adversarial Malware Binaries, EUSIPCO2018
ARMS RACE

Attacks
- L-BFGS method
- Fast Gradient Sign Method (FGSM)
- Deep Fool
- C&W attack
- Universal Perturbation

Defenses
- Gradient masking
- Distillation
- Label smoothing
- Adversarial training
- Robust training

Biggio et al. (2017), Security Evaluation of Pattern Classifiers under Attack

Rapidly growing research field

Adversary-aware machine learning
SOCCRATES PERSPECTIVE

- Majority of attacks and defenses have been proposed in the context of computer vision

- Our goal is to investigate their applicability to the security domain
  - Network- and log-based anomaly detection
  - Attack detection (classification)
  - Malware classification

- Attacks: efficient methods for crafting adversarial examples
- Defenses: increasing robustness of the models

- Credibility of the threats in real-world scenarios
  - Develop stronger attack models
Machine learning algorithms have been shown to be vulnerable to the adversarial examples.

Adversarial examples can be crafted for most of the machine learning methods. They can be used to evade classifiers and raise potential security and safety threats.

Future Outlook:
- Adversary-aware machine learning and pro-active defenses
- Build stronger attacker models
- Keep the security tools up-to-date and maintain their robustness against the latest attacks

CONCLUSIONS
THANK YOU!

Ewa Piatkowska
ewa.piatkowska@ait.ac.at