Resilient Machine Learning in Adversarial Environments

Alina Oprea
Associate Professor
Northeastern University

November 5, 2019
Problem space

• **Space**: Adversarial Machine Learning (study security of machine learning algorithms under various attacks)

• **Problem**: Need to test resilience of ML and AI algorithms in critical applications (cyber security, connected cars) and design robust ML methods

• **Solution**: New optimization-based testing time and training-time attacks against ML classifiers; resilient linear models

• **Results**: Most ML algorithms are vulnerable; resilient ML models are needed

• **TRL**: High for attacks; low for defenses

Alina Oprea
Associate Professor, Northeastern University
a.oprea@northeastern.edu
AI in Critical Applications

• AI has potential in critical applications
  – Cyber security: intelligent defense algorithms
  – Connected cars: assist and warn drivers of safety issues
  – Healthcare: assist doctors in diagnosis and treatment

• ...But AI could become a target of attack
  – Traditional ML and deep Learning are not resilient to adversarial attacks
  – Consider entire AI lifecycle from training to testing
  – Many critical real-world applications are vulnerable
  – New adversarially-resilient algorithms are needed!
## Adversarial Machine Learning: Taxonomy

<table>
<thead>
<tr>
<th>Learning stage</th>
<th>Targeted</th>
<th>Availability</th>
<th>Privacy</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Training</strong></td>
<td>Targeted Poisoning, Backdoor Trojan Attacks</td>
<td>Poisoning Availability, Model Poisoning</td>
<td>-</td>
</tr>
<tr>
<td><strong>Testing</strong></td>
<td>Evasion Attacks, Adversarial Examples</td>
<td>-</td>
<td>Membership Inference, Model Extraction</td>
</tr>
</tbody>
</table>
Evasion Attacks

**Evasion attack**: attack against ML at testing time

**Implications**
- Small (imperceptible) modification at testing time changes the classification
- Attacks are easy to mount and hard to detect

Adversarial example
Most evasion attacks done in the context of image classification

Example: Malicious connection classifier (features aggregated by port)

Challenge: Attacks designed for continuous domains do not result in feasible adversarial examples in discrete domains
Adversarial Framework in Discrete Domains

• General optimization framework for adversarial attacks in discrete domains
  – Respect *mathematical dependencies* (e.g., aggregated feature statistics)
  – Respect *physical-world constraints* (e.g., min and max packet size)
• Threat model
  – *Insert* realistic network connections (e.g., Bro conn events)
• Considered two cyber security applications
  – Public dataset for malicious network traffic classification
  – Enterprise dataset for malicious domain classification

• Evasion attacks can be easily mounted in discrete domains
• General framework applicable to multiple applications
How Effective are Evasion Attacks in Security?

- Malicious connection classifier can be easily attacked by inserting a small number of connections (12 new Bro logs)
- Significant degradation of ML classifiers under attack
Adversarial Example in Connected Cars

Original Image; steering angle = -4.25

Adversarial Image; steering angle = -2.25

• Convolutional Neural Networks used for steering angle prediction can be easily attacked
• Considered both classification and regression prediction tasks
Poisoning Availability Attacks

• **Attacker Objective**: Degrade model predictions
• **Capability**: Insert poisoning points in training

• Linear regression can be easily poisoned at training time
• Can train a resilient regression model by using our defense
Resilient Linear Regression

- Given dataset on $n$ points and $\alpha n$ attack points, find best model on $n$ of $(1 + \alpha)n$ points
- If $w, b$ are known, find points with smallest residual
- But $w, b$ and true data distribution are unknown!

- TRIM: robust optimization defense
- Solve a trimmed optimization problem using a subset of points
- Provable guarantees of worst-case attack impact
Network and Distributed System Security (NDS2) Lab

• Machine learning and AI for cybersecurity
  – Threat detection
    • [Yen et al. 13], [Yen et al. 14], [Oprea et al. 15], [Li and Oprea 16], [Buyukkayhan et al. 17], [Oprea et al. 18], [Duan et al. 18], [Ongun et al. 19]
  – Collaborative enterprise defense: Talha Ongun (PhD student), Oliver Spohngellert (MS student), Simona Boboila (Research Scientist)
  – IoT security: Talha Ongun
  – AI for cyber security games: Lisa Oakley (RS), Giorgio Severi (PhD student)
• Adversarial machine learning and AI
  – Poisoning attacks and defenses [Liu et al. 17], [Jagielski et al. 18], [Demontis et al. 19]: Matthew Jagielski (PhD student); Niklas Pousette Harger; Ewen Wang (undergraduate)
  – Evasion attacks for cyber security and connected cars [Chernikova et al. 19], [Chernikova and Oprea 19]: Alesia Chernikova (PhD student)
  – Privacy and fairness [Jagielski et al. 19]: Matthew Jagielski; Alesia Chernikova
Acknowledgements

Contact Information
Alina Oprea
a.oprea@northeastern.edu

Northeastern University
Cybersecurity & Privacy Institute
Evasion Attacks

• [Szegedy et al. 13] Intriguing properties of neural networks
• [Biggio et al. 13] Evasion Attacks against Machine Learning at Test Time
• [Goodfellow et al. 14] Explaining and Harnessing Adversarial Examples
• [Carlini, Wagner 17] Towards Evaluating the Robustness of Neural Networks
• [Madry et al. 17] Towards Deep Learning Models Resistant to Adversarial Attacks
• [Kannan et al. 18] Adversarial Logit Pairing
• ...

Adversarial example
Evasion Attacks For Neural Networks

Input: Images represented as feature vectors

Optimization Formulation

Given input $x$
Find adversarial example

$$x' = x + \delta$$

$$\min_{\delta} c||\delta||^2_2 + Z_t(x + \delta)$$

Min distance  Change class

• Existing attacks: [Carlini and Wagner 2017], [Biggio et al. 2013], [Madry et al. 2018]

• Challenge: Attacks designed for continuous domains do not result in feasible adversarial examples in cyber security (feature extraction layer)
Evasion Attacks for Security

Challenge
• Attacks designed for continuous domains do not result in feasible adversarial examples

Solution
• New iterative attack algorithm taking into account feature constraints
Adversarial Framework for Discrete Domains

Input: adversarial objective $A(x)$
- original point $x_0$; target class $t$
- learning rate $\alpha$; $D$ dependent feature set

Repeat until stopping condition:
1. $i \leftarrow \text{argmax} \nabla_x A(x)$ // Feature of max gradient
2. if $i \in D$
   1. $x_r \leftarrow \text{Find} \_\text{Representative}(i)$ // Find family representative
   2. $x_r \leftarrow \Pi(x_r - \alpha \nabla_{x_r} A(x))$ // Gradient update of representative feature
   3. $\text{Update} \_\text{Dependecies}(i)$ // Update all dependent features
3. else
   1. $x_i \leftarrow \Pi(x_i - \alpha \nabla_{x_i} A(x))$ // Gradient update for feature $i$

if $C(x) = t$ return $x$ // Found adversarial example
Evasion Attack for Malicious Connection Classifier

<table>
<thead>
<tr>
<th>Time</th>
<th>Src IP</th>
<th>Dst IP</th>
<th>Prot.</th>
<th>Port</th>
<th>Sent bytes</th>
<th>Recv. bytes</th>
<th>Sent packets</th>
<th>Recv. packets</th>
<th>Duration</th>
</tr>
</thead>
<tbody>
<tr>
<td>9:00:00</td>
<td>147.32.84.59</td>
<td>77.75.72.57</td>
<td>TCP</td>
<td>80</td>
<td>1065</td>
<td>5817</td>
<td>10</td>
<td>11</td>
<td>5.37</td>
</tr>
<tr>
<td>9:00:05</td>
<td>147.32.84.59</td>
<td>87.240.134.159</td>
<td>TCP</td>
<td>80</td>
<td>950</td>
<td>340</td>
<td>7</td>
<td>5</td>
<td>25.25</td>
</tr>
<tr>
<td>9:00:12</td>
<td>147.32.84.59</td>
<td>77.75.77.9</td>
<td>TCP</td>
<td>80</td>
<td>1256</td>
<td>422</td>
<td>5</td>
<td>5</td>
<td>0.0048</td>
</tr>
<tr>
<td>9:00:20</td>
<td>147.32.84.165</td>
<td>209.85.148.147</td>
<td>TCP</td>
<td>443</td>
<td>112404</td>
<td>0</td>
<td>87</td>
<td>0</td>
<td>432</td>
</tr>
</tbody>
</table>

- **Family**: all features defined per port
- **Attack**: Insert TCP or UDP connections on the determined port
- **Representative features**: number of packets in a connection
- **Dependent features**: sent bytes, duration
  - Respect physical constraints on network

Raw Bro logs:

1. Insert TCP or UDP connections on the determined port.
How Effective are Evasion Attacks in Security?

- **Dataset**: CTU-13, Neris botnet
  - 194K benign, 3869 malicious
- **Features**: 756 on 17 ports
- **Model**: Feed-forward neural network (3 layers), F1: 0.96

- **Baseline 1**
  - Features selected at random
- **Baseline 2**
  - Features and values selected at random
How Effective are Evasion Attacks in Security?

Significant degradation under attack

Malicious connection classifier

Malicious domain classifier
Evasion Attacks in Connected Cars

- Udacity challenge 2: Predict the steering angle from camera images, 2014
- Actions
  - **Turn left** (negative steering angle below threshold -T)
  - **Turn right** (positive steering angle above threshold T)
  - **Straight** (steering angle in [-T,T])
- The full dataset has 33,608 images and steering angle values (70GB of data)

Predict direction: Straight, Left, Right
Predict steering angle


*Are Self-Driving Cars Secure? Evasion Attacks against Deep Neural Networks for Self-Driving Cars.*

CNN for Direction Prediction

- Two CNN architectures: 25 million and 467 million parameters
Evasion Attack against Regression

- First evasion attack for CNNs for regression task (predict steering angle)
- New objective function
  - Minimize adversarial perturbation
  - Maximize the square residuals (difference between the predicted and true response)

\[
\min_{\delta} c ||\delta||^2_2 - g(x + \delta, y)
\]
\[
\text{such that } x + \delta \in [0,1]^d
\]
\[
g(x + \delta, y) = [F(x + \delta) - y]^2
\]

- 10% of adversarial images have MSE 20 times higher than legitimate images
- The maximum ratio of adversarial to legitimate MSE reaches 69
By changing only minimally the images (0.8 L2 perturbation), the attack has 100% accuracy!

Significant degradation of accuracy under attack from AUC = 1 to AUC = 0.62
Training-Time Attacks

• ML is trained by crowdsourcing data in many applications

  • Social networks
  • News articles
  • Tweets

  • Navigation systems
  • Face recognition
  • Mobile sensors

• Cannot fully trust training data!
Optimization Formulation

Given a training set $D$, find a set of poisoning data points $D_p$ that maximizes the adversary objective $A$ on validation set $D_{val}$ where corrupted model $\theta_p$ is learned by minimizing the loss $L$ on $D \cup D_p$

$$\argmax_{D_p} A(D_{val}, \theta_p) \text{ s.t. } \theta_p \in \argmin_{\theta} L(D \cup D_p, \theta)$$

First white-box attack for regression [Jagielski et al. 18]
- Determine optimal poisoning point $(x_c, y_c)$
- Optimize by both $x_c$ and $y_c$
Is It Really a Threat?

• Case study on healthcare dataset (predict Warfarin medicine dosage)
• At 20% poisoning rate
  – Modifies 75% of patients’ dosages by 93.49% for LASSO
  – Modifies 10% of patients’ dosages by a factor of 4.59 for Ridge
• At 8% poisoning rate
  – Modifies 50% of the patients’ dosages by 75.06%

<table>
<thead>
<tr>
<th>Quantile</th>
<th>Initial Dosage</th>
<th>Ridge Difference</th>
<th>LASSO Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>15.5 mg/wk</td>
<td>31.54%</td>
<td>37.20%</td>
</tr>
<tr>
<td>0.25</td>
<td>21 mg/wk</td>
<td>87.50%</td>
<td>93.49%</td>
</tr>
<tr>
<td>0.5</td>
<td>30 mg/wk</td>
<td>150.99%</td>
<td>139.31%</td>
</tr>
<tr>
<td>0.75</td>
<td>41.53 mg/wk</td>
<td>274.18%</td>
<td>224.08%</td>
</tr>
<tr>
<td>0.9</td>
<td>52.5 mg/wk</td>
<td>459.63%</td>
<td>358.89%</td>
</tr>
</tbody>
</table>
Poisoning Regression

• Improve existing attacks by a factor of 6.83

Predict loan rate with ridge regression
(L2 regularization)
Resilient Linear Regression

• Given dataset on $n$ points and $\alpha n$ attack points, find best model on $n$ of $(1 + \alpha)n$ points
• If $w, b$ are known, find points with smallest residual
• But $w, b$ and true data distribution are unknown!

$$\arg\min_{w,b,I} L(w, b, I) = \frac{1}{|I|} \sum_{i \in I} (f(x_i) - y_i)^2 + \lambda \Omega(w)$$

$N = (1 + \alpha)n, \quad I \subset [1, \ldots, N], \quad |I| = n$
References

• Evasion attacks

• Poisoning attacks
  – C. Liu, B. Li, Y. Vorobeychik, and A. Oprea. *Robust Linear Regression Against Training Data Poisoning*. In AISEC 2017

• Transferability of attacks