Threats to Machine Learning Applications

Mark Sherman
Director, Cybersecurity Foundations, CERT
August 18, 2020

Software Engineering Institute
Carnegie Mellon University
Pittsburgh, PA  15213
Carnegie Mellon Leads an Ecosystem of Innovation for Cybersecurity

CMU Campus – Global Research University

- Global research university known for its world-class, interdisciplinary programs in computer science, machine learning/artificial intelligence, engineering, business, arts, policy, and science
- Ranked #1 for Computer Science, #1 for Artificial Intelligence, #6 in Engineering (U.S. News and World Report)
- 1,442 total faculty and 130 research centers
- CyLab, CMU's security and privacy research institute, brings together experts from all schools across the university

CMU Software Engineering Institute (SEI)

- Founded in 1984 by the DoD as a Federally-Funded Research and Development Center (FFRDC) focused on software engineering
- Leader in software engineering, cybersecurity, and artificial intelligence research
- Established CERT in 1988
- About $145M annual funding (~$23M DoD Line)
- Critical to the DoD ability to acquire, develop, operate, and sustain software systems that are innovative, affordable, trustworthy, and enduring (CMU SEI Sponsoring Agreement)
CERT Division

Founded on a unique combination of experiential understanding of DoD missions, the cyber warfighter, the operational domain, and constantly changing technology

Adapts the best science to impact operational missions, increase the trustworthiness of technology, and develop cyber talent

Partners with DoD, non-DoD agencies, and the private sector enable CERT to maintain technical depth, attract top talent, amplify DoD financial investment, reduce the risk to DoD missions, and scale the research

Strengthens the resilience of critical national functions, increases the cybersecurity and resilience of DoD systems and Defense Industrial Base, and develops the cyber capacity of allies and partners
Outline

Understanding the ML Attack Surface
Understanding Risks of Transfer Learning
Remedies and Limitations
Conventional Threats to Machine Learning
Developing a Machine Learning Application

Adapted from Joseph, Nelson, Rubinstein, Tygar; Adversarial Machine Learning, Cambridge University Press, 2019
Data Attacks – Selected Domain Subset

Adapted from Joseph, Nelson, Rubinstein, Tygar; Adversarial Machine Learning, Cambridge University Press, 2019
Data Attacks – Measurements

Adapted from Joseph, Nelson, Rubinstein, Tygar; Adversarial Machine Learning, Cambridge University Press, 2019
Algorithm Attacks – Feature Selection

Adapted from Joseph, Nelson, Rubinstein, Tygar; Adversarial Machine Learning, Cambridge University Press, 2019
Data Attacks – Features

Adapted from Joseph, Nelson, Rubinstein, Tygar; Adversarial Machine Learning, Cambridge University Press, 2019
Data Attacks – Training Data

Adapted from Joseph, Nelson, Rubinstein, Tygar; Adversarial Machine Learning, Cambridge University Press, 2019
Algorithm Attacks – Model Construction

Adapted from Joseph, Nelson, Rubinstein, Tygar; Adversarial Machine Learning, Cambridge University Press, 2019
Data Attacks – Model Testing Data

Adapted from Joseph, Nelson, Rubinstein, Tygar; Adversarial Machine Learning, Cambridge University Press, 2019
Data Attacks – Ground Truth

Adapted from Joseph, Nelson, Rubinstein, Tygar; Adversarial Machine Learning, Cambridge University Press, 2019
Algorithm Attacks – Model

Adapted from Joseph, Nelson, Rubinstein, Tygar; Adversarial Machine Learning, Cambridge University Press, 2019
Data Attack – Loss Measurements

Adapted from Joseph, Nelson, Rubinstein, Tygar; Adversarial Machine Learning, Cambridge University Press, 2019
Outline

Understanding the ML Attack Surface
Understanding Risks of Transfer Learning
Remedies and Limitations
Conventional Threats to Machine Learning
Deep Neural Network Structure

Trained Deep Neural Network

Overview of Transferring Learning

Bolun Wang, Yuanshun Yao, Bimal Viswanath, Haitao Zheng, Ben Y. Zhao; “With Great Training Comes Great Vulnerability: Practical Attacks Against Transfer Learning,” 27th USENIX Security Symposium; Aug 15-17, 2018; pg 1281
Overview of Transferring Learning

Bolun Wang, Yuanshun Yao, Bimal Viswanath, Haitao Zheng, Ben Y. Zhao; “With Great Training Comes Great Vulnerability: Practical Attacks Against Transfer Learning,” 27th USENIX Security Symposium; Aug 15-17, 2018; pg 1281
Overview of Transferring Learning

Bolun Wang, Yuanshun Yao, Bimal Viswanath, Haitao Zheng, Ben Y. Zhao; “With Great Training Comes Great Vulnerability: Practical Attacks Against Transfer Learning,” 27th USENIX Security Symposium; Aug 15-17, 2018; pg 1281
Deep Layer Feature Extraction

Used when domains are close

Pro: Cheap training; good accuracy

Con: Adversary has deep knowledge of teacher
    Easier to exfiltrate model
    Easier to create adversarial input

Bolun Wang, Yuanshun Yao, Bimal Viswanath, Haitao Zheng, Ben Y. Zhao; “With Great Training Comes Great Vulnerability: Practical Attacks Against Transfer Learning,” 27th USENIX Security Symposium; Aug 15-17, 2018; pg 1281
Full model fine tuning

Used when domains are not close

Pro: Better accuracy than deep layer feature extraction
Resilient to teacher-specific attacks

Con: Costly to train

Bolun Wang, Yuanshun Yao, Bimal Viswanath, Haitao Zheng, Ben Y. Zhao; “With Great Training Comes Great Vulnerability: Practical Attacks Against Transfer Learning,” 27th USENIX Security Symposium; Aug 15-17, 2018; pg 1281
Compromise choice

- Accuracy depends on relationship between student and teacher domains
- Better resiliency than deep, not as good as full
- More costly to train than deep, cheaper than full

Bolun Wang, Yuanshun Yao, Bimal Viswanath, Haitao Zheng, Ben Y. Zhao; “With Great Training Comes Great Vulnerability: Practical Attacks Against Transfer Learning,” 27th USENIX Security Symposium; Aug 15-17, 2018; pg 1281
Outline

Understanding the ML Attack Surface
Understanding Risks of Transfer Learning
Remedies and Limitations
Conventional Threats to Machine Learning
Creating Classifications

Feature Space → Deep Neural Net → Feature Space

Single input → Single classification
Adversarial Input

Feature Space \rightarrow \text{Deep Neural Net} \rightarrow \text{Feature Space}

Single input

"Inclusion Attack"

"Exclusion Attack"
Adding Resiliency

- Cutting off spikes mitigates undesired “inclusions”
- Enclosing spikes mitigates undesired “exclusions”
Training for resilience

Methods to improve model resiliency

• Add adversarial examples in training
• Train with larger domain subset
• Calculate convex hull of classification boundary
• Apply statistical robust regression

All of these methods trade resiliency for accuracy

• Adversarial examples are noisy
• Overfitting creates raggedy boundaries
• Concave boundaries could be legitimate – should be excluded
• Looser boundaries could be legitimate – should be included

Redundancy is an alternative strategy – at a cost
Outline

Understanding the ML Attack Surface
Understanding Risks of Transfer Learning
Remedies and Limitations
Conventional Threats to Machine Learning
Coding Hygiene

Any of the algorithms in creating the application or in the generated application could have coding weaknesses leading to vulnerabilities.

Mitigation: Good cyber hygiene

https://cve.mitre.org/cgi-bin/cvename.cgi?name=CVE-2020-5215
Software supply chain for assembled software

<table>
<thead>
<tr>
<th>Machine Learning Frameworks</th>
<th>Data Sources</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pandas</td>
<td>Kaggle</td>
</tr>
<tr>
<td>Numpy</td>
<td>UCI Machine Learning Repository</td>
</tr>
<tr>
<td>Scikit-learn</td>
<td>Find Datasets</td>
</tr>
<tr>
<td>Matplotlib</td>
<td>Data.gov</td>
</tr>
<tr>
<td>TensorFlow</td>
<td>xView</td>
</tr>
<tr>
<td>Keras</td>
<td>ImageNet</td>
</tr>
<tr>
<td>Seaborn</td>
<td>Google’s Open Images</td>
</tr>
<tr>
<td>Pytorch &amp; Torch</td>
<td></td>
</tr>
</tbody>
</table>
Machine learning system face training data supply challenges

Rich supplies of “deep fakes” are readily accessible

Poor detection of deep fakes

Cannot reliably verify that training data obtained through a supply chain

Preconfigured machine learning (i.e., teacher) systems provide a vehicle to distribute bad training data

Source: http://kaldir.vc.in.tum.de/faceforensics_benchmark/index.php (as of 9/25/19)
Reducing software supply chain risk factors

Software supply chain risk for a product needs to be reduced to acceptable level

- **Supplier Capability**: Supplier follows practices that reduce supply chain risks
- **Product Security**: Delivered or updated product is acceptably secure
- **Product Distribution**: Methods of transmitting the product to the purchaser guard against tampering
- **Operational Product Control**: Product is used in a secure manner

Denial of Service Attack

Remediation: Network hygiene
(https://us-cert.cisa.gov/ncas/tips/ST04-015)
Integration Points are Typically Weak

Machine learning applications are part of a system

New operating environments, i.e., interconnections between system parts, are a major cause of vulnerabilities

Extra-ML parts of the application are routes to ML attacks

Insider Threat

Easy vector for data attacks

Remediations:
• Organizational evaluation
• Organizational processes
• Tools
• Training

https://www.sei.cmu.edu/education-outreach/courses/course.cfm?coursecode=V26
"Fake News" and AI Untrustworthiness

People ultimately use output from ML systems
Reasoning from ML systems is generally opaque
Parties can amplify potential misgivings

“Through 2021, 80% of line of business (LOB) leaders will override business decisions made by AI,” Gartner survey*

Remediations:
• Technical: Improved explanations and expectations
• Social: Education and experience

Recognize: Machine Learning is Statistics

*Graham Peters, Alan D. Duncan, Gartner Group, “100 Data and Analytics Predictions Through 2024,” March 20, 2020, pg 4
Outline

Understanding the ML Attack Surface
Understanding Risks of Transfer Learning
Remedies and Limitations
Conventional Threats to Machine Learning
Ways to Engage with Us

- Download software and tools
- Explore research and capabilities
- Participate in education offerings
- Attend an event
- Search the digital library
- Read the SEI Year in Review
- Collaborate with the SEI on a new project

Software Engineering Institute
Carnegie Mellon University
4500 Fifth Avenue
Pittsburgh, PA 15213-3890
412-268-5800 - Phone
888-201-4479 - Toll-Free
412-268-5758 - Fax
info@sei.cmu.edu - Email
www.sei.cmu.edu - Web