Trusted AI in the Cyber Domain

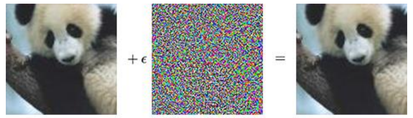
Presenter: Ritu Chadha

MOTIVATION

- Focus on Multi-Domain Operations increasingly relies on unmanned sensors and autonomy
- A lot of research is being conducted on Adversarial Machine Learning, but the general area of AI robustness is not well understood
- Even though ML attacks have been discussed mostly in the media domain, similar attacks in the cyber domain are just a matter of time

TYPES OF ADVERSARIAL MACHINE LEARNING (AML) ATTACKS

Small perturbations (imperceptible to humans) in input data can result in misclassification by ML algorithms



"panda" 57.7% confidence

"gibbon" 99.3% confidence

Insert "Trojans" into training data with specific labels designed to elicit specific outcomes

(Evtimov et al., UC Berkeley, 2017)

Stop signs with yellow "sticky notes" labeled "speed limit sign"

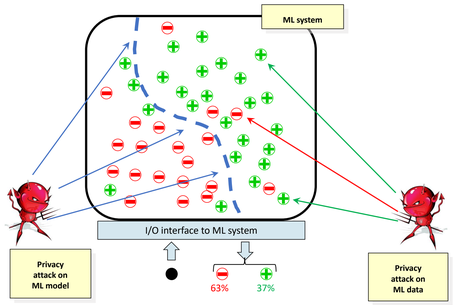
Lujo Bauer Mila Jovovich (87%)

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

Wearing a pair of eyeglass cutouts can fool facial recognition systems

Privacy Attacks:

- Model inversion: Gain access to sensitive data that was used to train the ML model
- Membership attacks: Determine whether or not a specific point was part of the training dataset analyzed to learn the parameter values of the model



Slide 1

ML PIPELINE ATTACK SURFACE

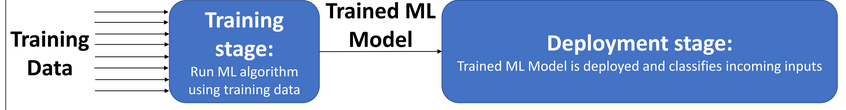
Training-time attack: Attacker manipulates data *labels* to perform **data poisoning attack** or **insert Trojans**

White box access: Attacker gains access to ML model code. Attacker uses ML Model to determine how to manipulate malicious inputs to make the ML model classify them as benign **(inference time attack)**

Black box access: Attacker is able to *probe* ML model (submit inputs, observe decisions)

Attacker constructs a replica of the model; attacker then uses this replica to determine how to manipulate malicious inputs to make the ML model classify them as benign **(inference time attack)**

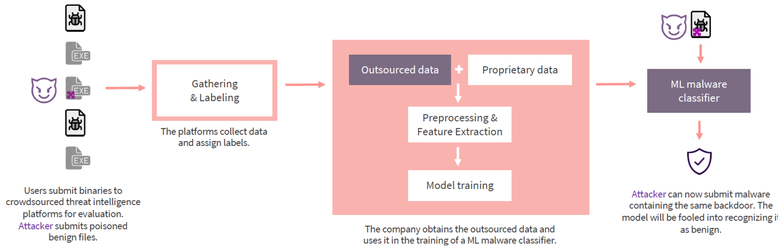
Attacker attempts to determine whether a specific data point was used to train the model **(privacy attack)**



The good news: AML attacks are much harder to perform on cyber models (compared to images)

- Need to *manipulate the raw data* so that *the features computed from the manipulated raw data (if any) will result in adversarial samples that will fool the ML model*, where the adversarial samples should satisfy the following constraints:
  - They should *not modify the semantics* of the original sample, and
  - They should result in an *erroneous classification* decision

POISONING THE TRAINING DATA



Attacker submits *watermarked benign binaries* that are *correctly labeled* but result in poor accuracy of ML malware classifier on malicious binaries with same watermark[2]

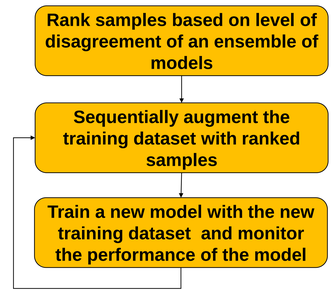
[1]Tianyu Gu, Brendan Dolan-Gavitt, and Siddharth Garg, "BadNets: Identifying Vulnerabilities in the Machine Learning Model Supply Chain", arXiv:1708.06733 [cs], August 2017

[2]Severi, Giorgio, Jim Meyer, Scott Coull, and Alina Oprea. "Explanation-Guided Backdoor Poisoning Attacks Against Malware Classifiers." In *30th USENIX Security Symposium (USENIX Security 21)*, 2021.

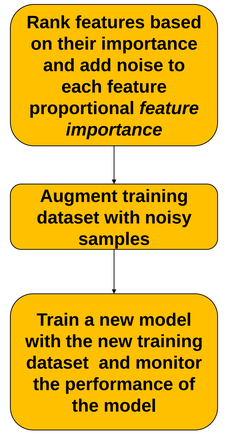
CASE STUDY

Defenses Against Poisoning Attacks on Malware Detectors

Data Sanitization using Nested Training: *Sequentially augment the training dataset* with samples based on disagreement of ensemble of models



Add noise to features of data samples in direct proportion to their feature importance



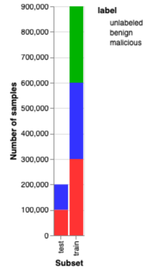
Why?

- Since the clean-label attack leverages regions in the feature space with the largest alignment towards goodware, the attack overfits the model to these features

➔ Adding noise to these most "important" features leads to higher loss for a poisoned than a clean model

DATASET

EMBER dataset: Includes features extracted from 1.1M Windows portable executable binary files



RESULTS

Results with 1% poisoned samples over 10 Runs[3]

Attack Configuration					
Model	Strategy	Watermark Size	Acc_{pois}	Acc_{clean}	Acc_{def}
EmberNN	Combined	32	3.42%	3.57%	93.2%
	MinPopulation	32	9.84%	100%	86.2%
	CountAbsSHAP	32	33.02%	100%	92.3%
LightGBM	Combined	8	9.84%	30.25%	72.2%
	MinPopulation	8	0.97%	58.65%	95.4%
	CountAbsSHAP	8	19.93%	75.87%	94.9%

[3]S. Venkatesan, H. Sikka, R. Izmailov, R. Chadha, A. Oprea and M. J. de Lucia, "Poisoning Attacks and Data Sanitization Mitigations for Machine Learning Models in Network Intrusion Detection Systems," 2021 IEEE Military Communications Conference (MILCOM), San Diego, CA, USA, 2021, pp. 874-879

Peraton | LABS