Demonstrating Data Poisoning Attacks on ML Models with Multi-Sensor Inputs

Shyam Maisuria Yuichi Ohsita Masayuki Murata

IST 大阪大学 OSAKA UNIVERSITY MURATA LAB. | Advanced Network Architecture Research Laboratory

1

Presentation Outline

- Background
- Poisoning Attack
- Experiment
- Results
- Conclusion

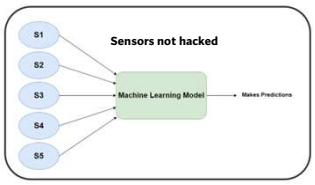
2

Background

3

Background[1]

- Many Machine Learning(ML) based systems uses multiple sensors
- It is possible that some sensors are hacked by the attacker
- The ML model can also be attacked from the hacked sensors
- Kurniawan et al demonstrated the possibility of this kind of attacks as an adversarial examples



[1] Ade Kurniawan, Yuichi Ohsita, and Masayuki Murata. Experiments on Adversarial Examples for Deep Learning Model Using Multimodal Sensors. Sensors, 22(22):8642, nov 2022.

4

Goal of Research

- **Objective:** Investigate data poisoning attacks and their profound impact on machine learning models.
- **What is Data Poisoning Attacks?**
  - Machine learning model is trained using compromised data.
- **Why Focus on Data Poisoning Attacks?**
  - **Significant Impact:** These attacks can severely compromise model integrity, making them a critical concern.
  - **Long-Term Consequences:** Once attacked, models may remain vulnerable over an extended period.
- **Our Research Aims To:**
  - Demonstrate the possibility of the data poisoning attack from a part of sensors

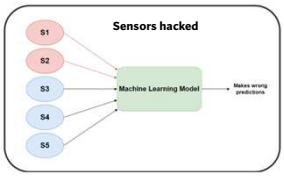
5

Poisoning Attack

6

Poisoning Attack Scenario

- Attacker's ability
  - An attacker hacked a part of sensors
  - Attacker can obtain the values of the hacked sensors
    - Attacker cannot obtain the values of the other sensors
  - Attacker can change the values of the hacked sensors before the values are sent to the model
  - We assume that the attacker has enough knowledge about the target model

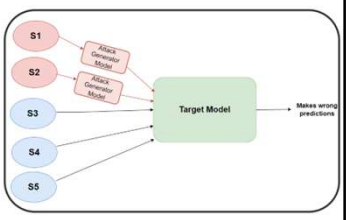


7/22

7

Generation of attacks

- The attacker installs an attacker generator model that changes the output of sensors.
- The attack generator model is trained in advance by using the knowledge of the attacker

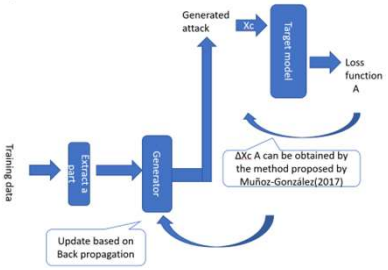


8/22

8

How to train the generator[2]

- **Input for Attack Generator Model**: Hacked sensor data
- **Output from Attack Generator Model**: Generated poisoned data
- **Training Objective**: Minimize loss function A to degrade target model accuracy



[2] Luis Mu˜noz-Gonz´alez, Battista Biggio, Ambra Demontis, Andrea Paudice, Vasin Wongrassamee, Emil C. Lupu, and Fabio Roli. Towards poisoning of deep learning algorithms with back-gradient optimization. AISec 2017 - Proc. 10th ACM Work. Artif. Intell. Secur. co-located with CCS 2017, pages 27–38, 2017.

9/22

9

Experiments

10

Secure Water Treatment Plant (Swat)[3]

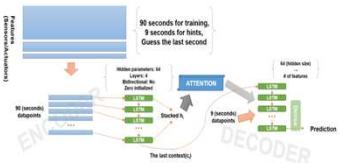
- Secure Water Treatment (Swat) is a water treatment site for cybersecurity research
- 11 days of continuous operation:
  - 7 days under normal operation
  - 4 days with attack scenarios
- 51 sensors and actuators
- Total of 6 different water treatment processors

[3]Jonathan Goh, Sridhar Adepu, Khurum Nazir Junejo, and Aditya Mathur. A dataset to support research in the design of secure water treatment systems. Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), 10242 LNCS(October):88–99, 2017.

11/22

11

Target Model[4]



Definition of Anomaly Score

- Anomaly Score was used as a metric

$$D_i = \sqrt{\sum_i (\hat{d}_{i,s} - d_{i,s})^4}$$

- Dt is prediction error

$$S = \min\left(\frac{H}{L \times R}, 1.0\right)$$

- S is anomaly score
- H and L are 90 and 20 percentile values of Dt
- S is greater than 0.3 anomaly is detected

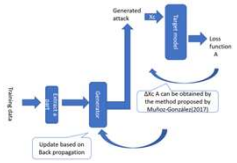
[4]Jonguk Kim, Jeong Han Yun, and Hyoung Chun Kim. Anomaly detection for industrial control systems using sequence-to-sequence neural networks. Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), 11980 LNCS:3–18, 2020.

12/22

12

Attackers' objective

- Objective function = weighted sum loss function
- $\mathcal{L}_T = \mathcal{A} \cdot \alpha + (1 - \alpha) \cdot \mathcal{L}_C$
- $\mathcal{L}_C$ = generator models objective $\mathcal{A}$= target models objective
- Alpha(α) controls the importance of each of the objective function
- Alpha(α) = low : prioritize evading detection
- Alpha(α) = high: prioritize effectiveness of attack

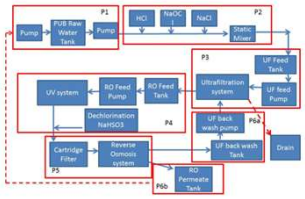


13/22

13

Target Model

- Each model is based on each process from Swat dataset
- We only focused on process 5
  - Which has total of 13 sensors and actuators
- Target model is trained using those 13 sensors and actuators
- Model accuracy: 71%
  - Found attacks: 5/7



14/22

14

Target Model – Features

35	AIT-501	Sensor	RO pH analyser; Measures HCl level.
36	AIT-502	Sensor	RO feed ORP analyser; Measures NaOCl level.
37	AIT-503	Sensor	RO feed conductivity analyser; Measures NaCl level.
38	AIT-504	Sensor	RO permeate conductivity analyser; Measures NaCl level.
39	FIT-501	Sensor	Flow meter; RO membrane inlet flow meter.
40	FIT-502	Sensor	Flow meter; RO Permeate flow meter.
41	FIT-503	Sensor	Flow meter; RO Reject flow meter.
42	FIT-504	Sensor	Flow meter; RO re-circulation flow meter.
43	P-501	Actuator	Pump; Pumps dechlorinated water to RO.
44	P-502 (backup)	Actuator	Pump; Pumps dechlorinated water to RO.
45	PIT-501	Sensor	Pressure meter; RO feed pressure.
46	PIT-502	Sensor	Pressure meter; RO permeate pressure.
47	PIT-503	Sensor	Pressure meter;RO reject pressure.

15/22

15

Hacked Sensors

35	AIT-501	Sensor	RO pH analyser; Measures HCl level.
36	AIT-502	Sensor	RO feed ORP analyser; Measures NaOCl level.
37	AIT-503	Sensor	RO feed conductivity analyser; Measures NaCl level.
38	AIT-504	Sensor	RO permeate conductivity analyser; Measures NaCl level.
39	FIT-501	Sensor	Flow meter; RO membrane inlet flow meter.
40	FIT-502	Sensor	Flow meter; RO Permeate flow meter.
41	FIT-503	Sensor	Flow meter; RO Reject flow meter.
42	FIT-504	Sensor	Flow meter; RO re-circulation flow meter.
43	P-501	Actuator	Pump; Pumps dechlorinated water to RO.
44	P-502 (backup)	Actuator	Pump; Pumps dechlorinated water to RO.
45	PIT-501	Sensor	Pressure meter; RO feed pressure.
46	PIT-502	Sensor	Pressure meter; RO permeate pressure.
47	PIT-503	Sensor	Pressure meter;RO reject pressure.

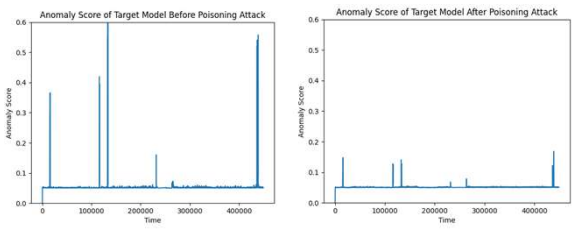
16/22

16

Results

17

Anomaly Scores of Target Model

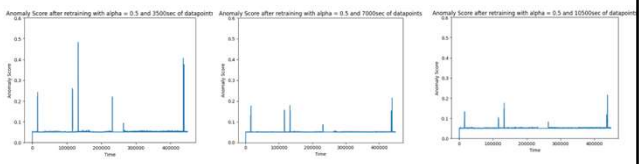


18/22

18

Anomaly Scores with varied data points

- Alpha = 0.5 is constant
- As number of datapoints increase anomaly scores decrease
- Shows poisoning attack is successful if we use more datapoints

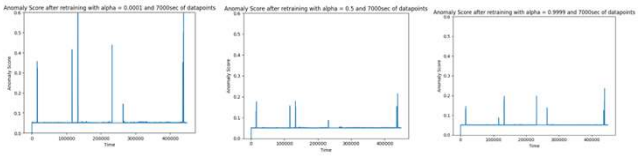


19/22

19

Anomaly Scores with varied alpha values

- Number of datapoints is constant at 7000sec
- As alpha value increases the anomaly score decreases
- Shows poisoning attack is successful when we increase the alpha value

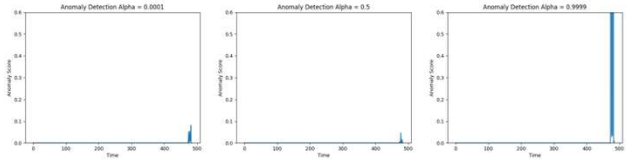


20/22

20

Check if Target Model detects Poisoned Attack

- Alpha values determine if attack is detected or not
- Higher alpha values = attack gets detected and poisoning attack is stronger
- Lower alpha values = avoid attack detection and poisoning attack is weaker



21/22

21

Conclusion

- Machine learning algorithms are **vulnerable to data poisoning** (compromising data collection), including Deep Learning systems.
- Target model is based on Swat anomaly detection
- Poisoning attack is generated using part of hacked sensors.
- Tested and attack was successful
- Next research objective is to work on countermeasures

22/22

22