2019
Research Experience for Undergraduates

Detection of Data Poisoning Attacks on Image Classification Models

Harsh Kachhadia
Co-Advisors: Dr. Amin Alipour
and Dr. Ioannis Kakadiaris

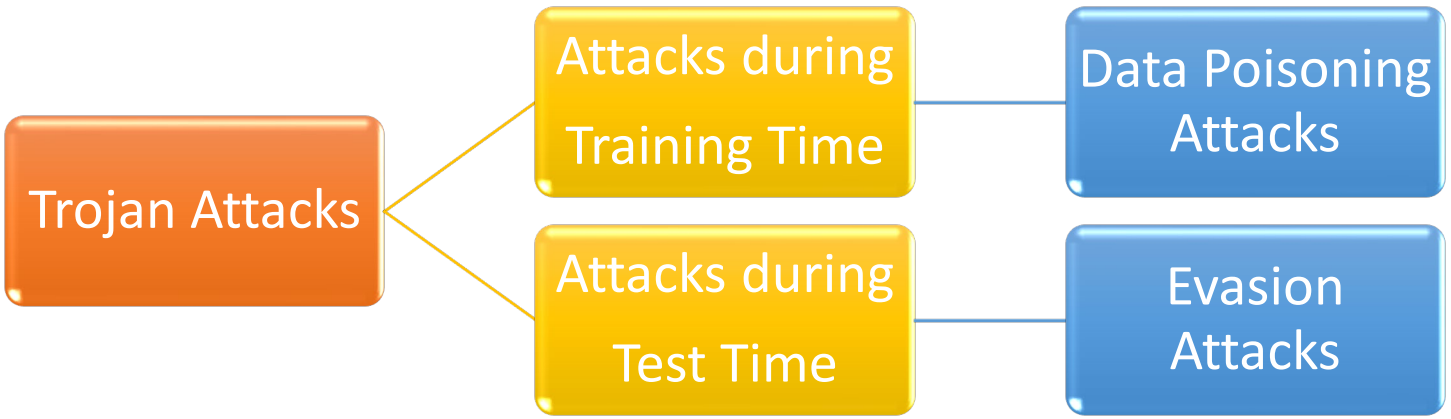
UNIVERSITY of HOUSTON

Motivation

- Deep Learning models, due to their amazing capabilities to solve difficult tasks such as Computer Vision, NLP, Game playing, etc. are being deployed into growing number of systems
- Talking about the rise of Deep Learning implementation in Computer Vision Field, it has made its place in various real-world applications such as Driverless Cars, Intelligent Doorbell, etc.
- But with the rise of its implementation, it has also started facing security and reliability issues
- Thus, before deployment of such systems in real-world, their robustness against such security attacks must be tested. Infact, Precautions must be taken to avoid such attacks in the first place

Background

- One such group of attacks are called Trojan Attacks; whose only intent is to make the model fail/malfunction during its test time



Background: Evasion Attacks

- These are the attacks that happen at the test time

How do they happen?

- First the attacker aims a targeted test image in the test image dataset and also aims for one of the classes that the model classifies images to.

- Attacker will then gain access to the test dataset and modify the targeted test image(in feature space) such that it will be misclassified into the targeted base class

- This can result into undesired outcome

Note: Having knowledge about the model and dataset distribution is necessary to carry out the attack successfully with/without access to the dataset.

Background: Data Poisoning Attacks

- These are the attacks that happen at the training time

How do they happen?

- Deep neural networks require large datasets for training and hyper-parameter tuning

- As a result, many practitioners turn to the web as a source for data, where one can automatically scrape large datasets with little human oversight

- While going for such automated data collection, an attacker injects maliciously crafted examples(poisoned data) into the training set in order to hijack the model and control its behaviour

- As a result, the trained model will be a compromised one. Attacker can easily use the model for fulfilling his own malicious tasks

Note: Having knowledge about the model and dataset distribution is necessary to carry out the attack successfully with/without access to the dataset

Goal

Establishing the baseline for detecting trojan attacks in image classification models.

Objectives

1. To create a data corpus for various data poisoning techniques

2. To train various deep learning models for image classification with and without data poisoning, let's call them regular and poisoned models

3. To compare the weights and activations of the regular and poisoned models

Expected Impact

- Compromised DNN models impact the operation of systems in the field and can endanger safety and security of users and communities

- This project will provide insights on how to ensure the integrity of training data in order to avoid data poisoning attacks

Deliverables

1. Datasets of various poisoning attempts
2. Source code for various neural network models
3. A final report

Tasks

(**Note:** Initially we will be working with MNIST & Keras.)

	Objective #
1. Know the Dataset. (Currently working on standardized image datasets) 2. Explore various Image Data Poisoning Techniques/Attacks 3. Find out how to implement them to poison a Dataset 4. Poison the Dataset	1
5. Implement an Image Classification Model on both regular and poisoned dataset	2
6. Compare and Observe the weights and activations of both the models using some tools such as Tensorboard or Pandas(Python Library) 7. Note down Observations and Derive Conclusions 8. Generate a Report based on observations and conclusions	3

Objective 1: Results

1. Explored the Labelling & Structure of MNIST Dataset

2. Explored Data Poisoning Techniques/Attacks

- ✓ Mislabelling Technique
- ✓ Making Random Changes in Image Pixel Values
- ✓ Change Orientation of Object in the Image
- ✓ Clean Label Attack for Transfer Learning
- ✓ Watermark Attack for End-to-End Learning
- ✓ Using GAN, to generate new Poisoned Face Dataset for Face Images

Objective 1: Results (2)

3. How to implement explored Data Poisoning Techniques

- ✓ Mislabelling Technique can be implemented by Targeted/Non-Targeted change in labels of images of dataset

- ✓ Making Random Changes in Image Pixel Values can be done my accessing pixel values present in numpy arrays using ‘for’ loops

4. Poison the Dataset

- ✓ Successfully poisoned MNIST Dataset using Mislabelling Technique

Objective 1: Remaining Work

- Explore more standard image datasets and some face datasets
- Explore how to implement the remaining listed poisoning techniques in python

Objective 2: Results

5. Implement an Image Classification Model on both regular and poisoned dataset

✓ Implemented a simple CNN Model in Keras on both poisoned and regular MNIST dataset

Objective 2: Remaining Work

- Implementation of better Image Classification models on datasets

Objective 3: Remaining Work

- Learn to use Tensorboard
- Observe the trained Image Classification models of MNIST dataset on Tensorboard
- Explore more options of Tensorboard Tool to have better observation of trained model on both poisoned and regular image datasets
- Find out, how to easily compare large data-frames of final weights of models trained on both regular and poisoned image datasets in pandas

Conclusions

Deep Learning Neural Networks along with their amazing capabilities also bring some loopholes, which if exploited by attackers, can result into disastrous result. It clearly points towards the rise in necessity to update our model with all kinds of latest defence techniques/mechanisms to prevent undesired outcomes before getting deployed in the real world

Acknowledgements

The REU project is sponsored by NSF under award NSF-1659755. Special thanks to the following UH offices for providing financial support to the project: Department of Computer Science; College of Natural Sciences and Mathematics; Dean of Graduate and Professional Studies; VP for Research; and the Provost's Office. The views and conclusions contained in this presentation are those of the author and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the sponsors.

UNIVERSITY of HOUSTON

Thank You

UNIVERSITY of HOUSTON