CS529 Security Analytics Fall 2023

Homework 3

Due: **November 30, 2023, 11:59 PM**

Descriptions & Instructions:

The goal of this homework is to help you become familiar with the implementation of the adversarial attacks/defenses and backdoor attacks covered in the lectures. You are provided with the Jupyter Notebook, **HW3.ipynb,** and you are required to complete the coding corresponding to the following questions. After completing the code segments, please ensure that the entire code in the notebook executes without any errors. The notebook walks you through generating adversarial images through targeted and non-targeted attacks. And it also introduces you to a very basic defense technique (binary thresholding). Given these introductions, you'll be asked to complete certain functions to implement these functionalities. The notebook also walks you through a backdoor attack. You will need to implement the attack from scratch, and evaluate its effectiveness.

Submit format:

You will compress your Jupyter Notebooks source code into a **zip file** called HW3.zip and submit them on **Brightspace**.

Then you need to submit a report including your solutions to the coding problems. You can choose to convert your Jupyter Notebook to a pdf report or take screenshots of your code and results. The report should be submitted on **Gradescope**.

Please mark your solutions to each question correctly while submitting the report on Gradescope.

Failure to follow the instructions will lead to a deduction in points!

Assignment 1: Targeted and Non-Targeted Adversarial Attacks, and Defenses [50 pt]

The goal of this question is to construct adversarial examples to attack a machine learning system (digit recognition). You will also create a simple defense and check whether the defense works.

To accomplish this assignment, please complete the following questions in HW3.ipynb:

1. Non-Targeted Attacks: A short introduction to such kinds of attacks is provided in the notebook. You have to complete the function that implements the idea of non-targeted attacks. [15 pts]

2. Targeted Attacks: An introduction on the same is given in the notebook, you have to complete the associated function to implement such an attack. [20 pts]

3. Simple defense against adversarial attacks: For this segment, you'll implement a simple technique as a defense strategy. Namely, you'll perform binary thresholding of the images and see whether this technique is effective. Complete the associated function(s) for this part. [15 pts]

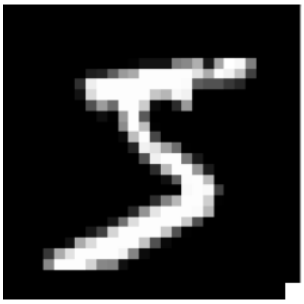
NOTE: You have to complete the segments marked explicitly, where you have to fill in the code, and there are some segments, where there is a None present, you have to fill these too, considering the comments provided in the notebook.

Assignment 2: Targeted Backdoor Attack [50 pt]

The goal of this question is to implement and launch the backdoor attack on a machine learning model, proposed by the paper - *BadNets: Identifying Vulnerabilities in the Machine Learning Model Supply Chain*.

To accomplish this assignment, first you need to **read and understand the paper,** then complete the following questions in HW3.ipynb:

1. Poison the dataset: [10 pts]
   - The trigger pattern is all-white 3*3 square patch at the bottom-right corner of an image.
     i. an example of a poisoned image, notice the patch at the bottom-right corner.



   - Target is label 0.

2. Training: [20 pts]
    - Train the Badnet model using the poisoned data.
    - Evaluate the accuracy on poisoned and clean testing data.

3. Visualization: [20 pts]
    - Report the final attack success rate and clean accuracy of your model.
    - Randomly select one clean image from the original testing set, plot the image, print the original prediction of the Badnet model that you trained. Then add the trigger to the previous image you choose, print the prediction of the Badnet model that you trained. It is expected that after adding the trigger, the prediction of the model should be the target that you previously selected (i.e. label 0).