FIT5230 Malicious AI

Week 3 Lab: Adversarial Machine Learning I

Objectives

This week's lab objectives are:

- To better understand the **concept** of adversarial machine learning & its **effects**
- To familiarise with different **types** of adversarial machine learning **attacks**
- To gain **hands-on** experience in adversarial machine learning by implementing some algorithms while leveraging on existing libraries

Please refer to the following Colabs and **make a copy** to your own Drive:
Week 3 Lab: Adversarial ML I - 2024 - Lab.ipynb for **Task 1 & 2**.
(Opt) Week 3 Advanced Lab: Adversarial ML I-2024-Lab.ipynb for **Task 3.**

Tasks

This week's lab tasks are designed to expose students to how adversarial machine learning techniques can be applied to actual datasets and to appreciate how they work to confuse the machine learning algorithms.

The lab tutor has written these codes by adapting older versions of GitHub/Colab that no longer run efficiently due to the obsoletion of libraries. The crucial sections requiring your implementation of attacks have been clearly marked for you to edit. In particular, you are to do Task 1 to implement simple black-box adversarial attacks; and then move on to Task 2 which will focus on a more complicated adversarial attack.

Task 1: Implement simple black box attacks

- Prepare a list of test images, for example, animals, vehicles, and objects.
- Implement Semantic Attack and Noise Attack
- Test the attack methods with your data

Task 2: Implement the Fast Gradient Sign Method (FGSM) & Fast Gradient Value Method (FGVM)

- Implement FGSM Attack
- Implement FGVM Attack based on FGSM Attack
- Test the attack methods with your data from Task 1

(Optional) Advanced Task 3: Generating backdoor adversarial attacks

- Implement Poisoning Attack
- Test the attack methods on MNIST Deep Learning Model

Reference: https://pytorch.org/tutorials/beginner/fgsm_tutorial.html