SCIT

School of Computing & Information Technology

CSIT375 – AI and Cybersecurity

Assignment 2

Due on Sunday, 08 Jun 2025 at 5:00pm

There are **3** tasks in this assignment which contribute to **20%** of the final marks. In addition, there is one optional task which has 3 **bonus marks**. These 3 bonus marks can be used to offset your lost marks in Task 1, Task 2 and Task 3. Your total marks will not exceed 20.

Upload this assignment folder to Google Drive. Detailed instructions can be found in **assignment2_CSIT375.ipynb**. Follow the instructions and run all the cells to complete tasks.

Task 1/3: Module Backdoor Attack (Total: 7 marks)

You will implement module backdoor attack by attaching a small module to the target model.

- **(5 marks)** Implement **train** in train_bad_module.py.
  - 5 marks if fooling rate ≥ 95% and decrease in accuracy ≤ 0.1%.
  - 3 marks if fooling rate ≥ 70% and decrease in accuracy ≤ 0.1%.
  - 2 marks if fooling rate ≥ 50% and decrease in accuracy ≤ 0.1%.
  - 0 marks otherwise.
- **(2 marks)** Briefly describe how you implement the module backdoor attack.
  - Write your answer in the corresponding text cell.

Task 2/3: Reverse Engineering (Total: 7 marks)

Given a trojaned model, you will reverse engineer the embedded trigger.

- **(5 marks)** Implement **reverse_trigger** in reverse_engineer.py.
  - 5 marks if fooling rate of the reversed trigger ≥ 95%.
  - 3 marks if fooling rate of the reversed trigger ≥ 70%.
  - 2 marks if fooling rate of the reversed trigger ≥ 50%.
  - 0 marks otherwise.

  Note that if your reversed trigger is significantly different from the original trigger, e.g., your reversed trigger looks like random noise, this means you fail to reverse the trigger and your marks will be manually changed to 0 for this coding part.
- **(2 marks)** Briefly describe how you reverse engineer the trigger.
  - Write your answer in the corresponding text cell.

Task 3/3: Data-free Adaptive Attack against DeepJudge (Total: 6 marks)

You will implement a data-free adaptive attack against DeepJudge.

- **(5 marks)** Implement **transform** in adaptive_attack.py.
    - 5 marks if decrease in accuracy < 0.5% and defeating DeepJudge.
    - 3 marks if decrease in accuracy < 1.5% and defeating DeepJudge.
    - 2 marks if decrease in accuracy < 3.0% and defeating DeepJudge.
    - 0 marks otherwise.
- **(1 mark)** Explain the rationale behind your adaptive attack.
    - Write your answer in the corresponding text cell.

Optional Task: Image Watermarks (Total: 3 bonus marks)

You will implement a deep watermarking technique that embeds subtle watermarks to clean images.

- **(2 marks)** Implement **train** in watermark.py.
    - 2 marks if TPR ≥ 95% and TNR ≥ 95% and FPR ≤ 5%.
    - 0 marks otherwise.

      It is acceptable that your watermarks are **slightly visible**. However, if your watermarks are **obvious noise**, this means you failed to generate **subtle** watermarks, and your marks will be manually changed to 0 for this coding part.
- **(1 mark)** Briefly describe how you implement the watermarks.
    - Write your answer in the corresponding text cell.

Submission

For submission, follow the instructions in the last cell in **assignment2_CSIT375.ipynb** to submit your work via Moodle.

The assessment must be your own work. If asked, you must be able to explain what you did and how you did it. Marks will be deducted if you cannot correctly explain these.

NOTE: The mark allocations shown above are merely a guide. Marks will be awarded based on the overall quality of your work. Marks may be deducted for other reasons, e.g., if your code is too messy, if you cannot correctly explain what you did or how you did it, etc.