UOW AUSTRALIA

SCIT

School of Computing & Information Technology

CSIT375 – AI and Cybersecurity

Assignment 1

Due on Sunday, 13 Apr 2025 at 5:00pm

There are **2** tasks in this assignment which contribute to **15%** of the final marks. In addition, there is one optional task which has 2 **bonus marks**. These 2 bonus marks can be used to offset your lost marks in Task 1 and Task 2. Your total marks will not exceed 15.

Upload this assignment folder to Google Drive. Detailed instructions can be found in **assignment1_CSIT375.ipynb**. Follow the instructions and run all the cells to complete tasks.

Task 1/2: Grey-box Adversarial Examples (Total: 8 marks)

You will implement grey-box targeted adversarial examples to fool a pretrained target model. The L_∞ norm of adversarial perturbations are required to be ≤ 0.04.

- **(6 marks)** Implement **generate_attack** in greybox_attack.py.
  - 6 marks if fooling rate $\geq 80\%$.
  - 4 marks if fooling rate $\geq 60\%$.
  - 2 marks if fooling rate $\geq 40\%$.
  - 1 mark if fooling rate $\geq 20\%$.
  - 0 marks otherwise.
- **(2 marks)** Briefly describe how you implemented the grey-box adversarial examples.
  - Write your answer in the corresponding text cell.

Task 2/2: Universal Adversarial Perturbations (Total: 7 marks)

You will implement targeted universal adversarial perturbations (UAPs) to fool a pretrained target model. The L_∞ norm of adversarial perturbations are required to be ≤ 0.06.

- **(5 marks)** Implement **generate_UAPs** in universal_attack.py.
  - 5 marks if fooling rate $\geq 90\%$.
  - 3 marks if fooling rate $\geq 70\%$.
  - 1 mark if fooling rate $\geq 50\%$.
  - 0 marks otherwise.
- **(2 marks)** Briefly describe how you implemented the UAPs.
  - Write your answer in the corresponding text cell.

Optional Task: Adaptive Attack (Total: 2 bonus marks)

You will implement white-box adaptive attack to bypass a stochastic defence. The L_∞ norm of adversarial perturbations are required to be ≤ 0.04.

- **(1 mark)** Implement **generate_attack** in adaptive.py.
  - 1 mark if fooling rate $\geq 90\%$.
  - 0 marks otherwise.
- **(1 mark)** Briefly describe how you implemented the adaptive attack.
  - Write your answer in the corresponding text cell.

Submission

For submission, follow the instructions in the last cell in **assignment1_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.

