CSE 40685/60685: Machine Learning for Embedded Systems

Lab 04 – Adversarial attacks on Raspberry PI

If you need any help, please feel free to reach out to Dr. Jun Xia at jxia4@nd.edu

Overview and objectives:

- In this lab, you will learn how to generate adversarial patches against a trained neural network.
- You will learn how to use already generated patches to mislead a neural network on the Raspberry Pi.

Deadlines and deliverables:

- This is a two-week lab and you are expected to complete it before the beginning of the class on Thursday 3/20.
- This lab assignment is individual. You may follow the tutorial attached to this lab.
- You need to upload a lab report (no more than two pages) by the deadline, which should include: 1) Your understanding of the code on **how to generate adversarial patches**; 2) Try to generate a few patches and identify the most effective one. Explain how well it works (including its size, and if it can trick the neural network when attached to different objects beyond humans). Discuss how different patch designs affect the effectiveness. Discuss how the size of the patch affects the effectiveness. Why? (**Make sure to include the most effective patch you made in the report)**

Specific Tasks:

- Follow the attached tutorial to generate your own patch.
- Evaluate the effectiveness of your patch on the performance of yolov2.

References:

[1] Thys S, Van Ranst W, Goedemé T. Fooling automated surveillance cameras: adversarial patches to attack person detection[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops. 2019.

[2] https://www.youtube.com/watch?app=desktop&v=MIbFvK2S9g8&embeds_euri=https%3A%2F%2Fhabr.com%2Fru%2Fcompany%2Fjetinfosystems%2Fblog%2F449216%2F&feature=emb_imp_woyt.