7/20 - 7/26 Progress Report

1. **Reading List**
    a. Rough Reading
        i. Tramèr, Florian, et al. "Ensemble adversarial training: Attacks and defenses." *arXiv preprint arXiv:1705.07204* (2017).
        ii. Feinman, Reuben, et al. "Detecting adversarial samples from artifacts." *arXiv preprint arXiv:1703.00410* (2017).
        iii. Xu, Weilin, David Evans, and Yanjun Qi. "Feature squeezing: Detecting adversarial examples in deep neural networks." *arXiv preprint arXiv:1704.01155* (2017).

2. **Ideas and Deliverables**
    a. Generating TGSM images and evaluate Saliency (started a few days ago)
        i. Having successfully generated and evaluated the saliency maps of FGSM images, I am now working on TGSM images. This allows the image to be more vigorously classified as another class, ultimately having a stronger impact in the saliency results. While prior FGSM definitely made an impact disrupting the CNN, the results were not significant and only caused a small shift in the output.
    b. Explore Robust Preprocessing Filters (recently started)
        i. This is a potential idea for my defence which is described in more detail in section 4a. I have not really started this part yet; however it is a potential concept and I am reading some papers and gathering information about this for now.
3. **Deliverables**
    a. Aside from the original FGSM and Saliency Map generation. I am almost done with the TGSM version. I will keep updated on the results of the TGSM version on GitHub and see if there is a potential project idea from this direction.

4. **Plan**
    a. Most of the baseline code is implemented and the preliminary results are documented. The immediate next step is to start implementing the project idea. While this part is still not set in stone, I have came up with a potential idea on defending against adversarial attacks. Most prior defences fall into 2 categories.
        i. Detecting adversarial images before feeding it into the CNN and ignoring them if they are classified as an attacked image. This may not be practical all the time. For example, an autonomous vehicle ignoring images will not allow it to process what is going on, proving hazardous on the road.
        ii. Training the model with adversarial images. This can take a lot of time due to the fact you have to retrain your model. Furthermore, when new attacks come out, the model have to be continually updated.

    While creating a new defence mechanism may be hard and out of the scope of the project, I am hoping to attack a subset of the problem. Results from my FGSM/Saliency Map results shows that adversarial images can cause a shift where the model focuses on when there are a lot of objects. For example, the initial saliency map of the outdoor terrace focused on the whole picture. After adding an FGSM attack on it, the model shifted focus to small specific areas, such as the bench. The correct classification of the model should be looking at the whole image not specific parts. Hence, I am hoping to create a preprocessing image filter for these attacked images that can help the model be more robust to changing its field of focus from the whole image to specific areas.