# 7/6 - 7/12 Progress Report

1. **Reading List**
    a. Rough Reading
        i. Ross, Andrew Slavin, and Finale Doshi-Velez. "Improving the adversarial robustness and interpretability of deep neural networks by regularizing their input gradients." *arXiv preprint arXiv:1711.09404* (2017).
        ii. Smilkov, Daniel, et al. "Smoothgrad: removing noise by adding noise." *arXiv preprint arXiv:1706.03825* (2017).
    b. Intensive Reading
        i. Goodfellow, Ian J., Jonathon Shlens, and Christian Szegedy. "Explaining and harnessing adversarial examples (2014)." *arXiv preprint arXiv:1412.6572*.
2. **Ideas and Deliverables**
    a. Generating FGSM images (started a week ago)
        i. The goal is to be able to have code that can generate the corresponding FGSM image of a given input.
    b. Comparing Saliency map of Original and FGSM images (recently started)
        i. The goal is to get the saliency maps of any input image as well as the saliency map of the corresponding FGSM image. This way we can compare how the model evaluates the pictures differently
3. **Deliverables**
    a. Right now I am able to generate the FGSM adversarial image of correctly sized handwritten numerical images. Furthermore, I am able to compute the saliency of any of these images. This means I am able to compute the saliency map of an original MNIST image as well as the saliency map of the FGSM image. However I am unable to do it on other datasets because I have yet to incorporate other models in the code. This is something I am working on.
4. **Plan**
    a. The immediate plan following this week is to be able to use SmoothGrad on the same images as where the FGSM images are generated out of.

The challenge is the current code uses different models hence we can not do that. By the end of the week I want to incorporate inception V3 in the FGSM generation code so I can generate FGSM images directly attacking inception V3. This will allow us to use SmoothGrad on the new FGSM generated images.

b. The long term plan is to figure out what we are going to do with our findings. Depending on what results we get from part *a*, we need to figure out what our contribution will be. Some tentative ideas are creating a new attack that better change the focus of the image as FGSM’s main contribution is to trick the model to make a wrong prediction. Another idea is to see how we can improve the model to not be so prone to having the model shift image focus. The last idea, which is a lot simpler, is just to learn about why the image focus is changing from simple FGSM attacks.

5. **Misc**

a. RIght now I am utilizing two code base as my baseline. One of which is the SmoothGrad code and the other being the Adversarial Attack code. The issue is that both of them works on different dataset/models. I think it is important soon to be able to modify them so I can choose one dataset and model to work with and still have access to the same functionality.