6/30 - 7/5 Weekly Progress

Goal From the Previous Week:

From the brainstorm progress of last week, my main goal for this week was to read the Adversarial Paper (Ross, et al. 2017) more in depth and figure out possible ways to incorporate the previous paper about sharpening saliency maps, SmoothGrad (Smilkov, et al. 2017).

Work Accomplished:

I have spent the week going over the adversarial paper in detail and figuring out what I can utilize in my project. In short, this paper analyzes different types of adversarial attacks (FGSM, TGSM, and JSMA) on three different datasets (MNIST, notMNIST, and SVHN). They have 5 different types of trained models with the baseline being a “5x5x32 and 5x5x64 convolutional layers followed by 2x2 max pooling and a 1024-unit fully connected layer with batch-normalization after all convolutions and both batch-normalization and dropout on the fully-connected layer.” This is implemented in Tensorflow and trained using Adam (Kingma and Ba 2014). Other models consists of Distilled CNN, AdvTrain CNN, Insensitive CNN, and DoubleBack CNN.

While there are many concepts in this paper, below I have simplified things by listing the techniques in this paper I believe can be utilized in my project. This is a tentative list and bound to change as I finalize a project topic.

Adversarial Attack:

I have chosen to focus mainly on Fast Gradient Sign Method “FGSM” (Goodfellow, et al. 2014) as the adversarial attack I will use. This is one of the first method in the field of adversarial attacks and many new adversarial attacks are developed based off of this method. While other techniques may have better results, I believe this will work better for my project due to the simplicity in its algorithm and the findings we obtain from this method usually reflects the similar behaviors from other methods.

Dataset:

The dataset is likely to change based off of what my final project topic is; however, that is not a big deal and easy to adjust to. With that said, given the choice, I will like to build my project around using the MNIST dataset as there are a lot of adversarial papers using this dataset. This will allow me to have more resources as the project continues. But again, this is not an important factor as of now.

Model:

While this project compares 5 different models, I find that a little excessive. If my project ends up comparing different models, I have narrowed it to 3 useful ones from this project.

- First, we will have the normal CNN with the architecture mentioned in the first paragraph of this section.
- Second, we will have AdvTrain CNN. In short, this CNN is trained using many different adversarial images as well. This allows the model to be more robust.
- Third, we have DoubleBack CNN which is trained with gradient regularization on its loss function.

I have omitted Distilled CNN due to its poor performance in this paper. I have also omitted Insensitive CNN because it is similar to DoubleBack CNN (both gradient-regularized models), just performs a little worse.

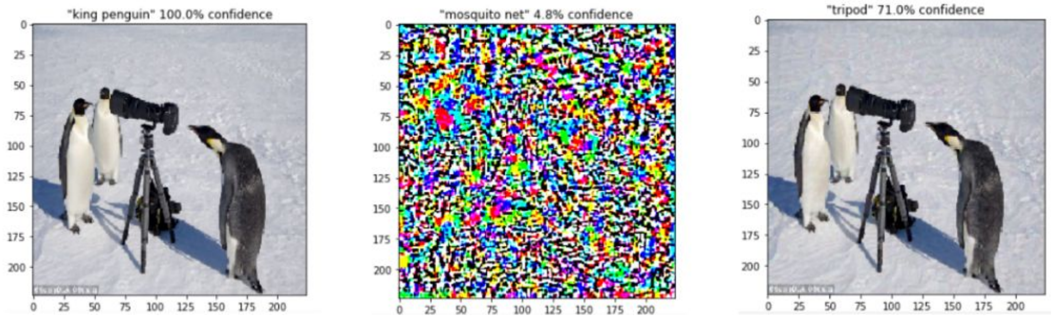
Current Ideas:

After reading the paper, I have given thought to some potential project idea. Hopefully we can come up with one to finalize on during the meeting.

- Adversarial attacks can happen in different ways. For example, putting a sticker on a stop sign can cause a self driving car to misclassify a stop sign as a speed limit sign (Schneier). We can explore how some of the successful models this paper used: AdvTraining, Doubleback, and etc. works against real world adversarial attacks in self driving car settings.

- Adding adversarial attacks like FGSM can not only change the prediction of a model, but also change the object it focuses on as shown in the image below. I can use saliency maps and SmoothGrad to investigate and learn why the saliency map changes as we increase the epsion in FGSM.

$$x \quad + \quad \varepsilon \, \text{sign}(\nabla_x L(\theta, x, y)) \quad = \quad x_{\text{adv}}$$


"king penguin" 100.0% confidence
"mosquito net" 4.8% confidence
"tripod" 71.0% confidence
0
25
50
75
100
125
150
175
200

- Design and implement a new model and test how it compares with the previous ones in this paper, hopefully improving it in some aspect.

Goals for Next Week:

I would like to finalize the topic for my project defense during our meeting. For the next week I would like to have a baseline of where to start deployed on github as well as a schedule markup on goals for the following weeks.