HW 05 Paper 02
Riley Densley

Deep Nets are Easily Fooled

This paper goes into how DNN's recognize images. They go into how the DNN can recognize both real and unrecognizable images with great certainty and why the DNN does this. Specifically how human vision recognize and image vs how a DNN classifies and image.

They introduce this idea of generating images with evolution. This is where they take images and slowly manipulate them until the DNN believes it is an actual image. This was a bit difficult for me to understand. What I think they did was feed the DNN a bunch of randomly generated images. Then those images that scored the best through the DNN were then used in the next generation. This continued until the DNN thought all images were natural images where in reality they were just a multitude of random pixels. They also talked about doing similar alterations of natural images and continually changing them until they were no longer recognizable by humans but were still being classified by the DNN.

It is interesting to me to see that the images were all just noise. They discuss this in the paper as well. They said that they didn't start this project in hopes of deceiving a DNN this way. The thought was that the EA would have produced recognizable images or images with a low confidence level. But the opposite happened. Although from what I understood the only reason that it was that confident was because the class was given to the DNN.

Beyond that I didn't get much more. They went into the need of having a good neural network to build images off of. They also talked about how they manipulated the images to get the best possibility of recognition from the neural network.