This paper was impressive because it shows that neural networks are susceptible to attacks that humans would have no problem avoiding. The main problem with the structure of networks is that they must make a choice. They are given a set of input data and based on the status of the weight and biases of the network at that moment, its answer is deterministic. At this point, general AI cannot work because there is not enough data of everything single thing nor a network complex enough to be trained on everything. Additionally, once a network is trained, and a new output is desired, the system cannot be altered. The network has to be retrained entirely with the new architecture and super parameters. This is why the network is not capable of changing itself or evolving and enhancing its capabilities to identify new objects on the fly. This is only what humans and other forms of life are capable of.

This is not to say that artificial networks are useful, actually the contrary. However, the limits of networks are apparent. They can excel in mundane tasks of identifying known objects, they can beat people in games once they played millions of games against itself. But without data, or a meager set of 10-20 examples will never be able to train a network like a human would be able to learn from. The network is a mathematical operation which takes numerical data, preforms various mathematical operations on it, then returns an answer. The heart of a network is no more complex than that the math done in a linear algebra course. But since the matrixes are much more large and complex than humans can purposely identify, we give it to a network to find arbitrary relationships in the data. We design the networks with logic and structure and then the network preforms calculations and slightly updates its own weight and biases (but not structure).

I found it interesting at the end of section 3.4, the network was harder to fool with cats and dogs. This was attributed to the larger number of cats and dogs that were in the system. The paper states, “the size of the dataset of cats and dogs it has been trained on is larger than for other categories, meaning is less overfit, and thus more difficult to fool.” I find this interesting because overfitting, until now in this class, has not been talked about much. However, the holy grail of networks is on that has been trained on a large dataset, it has a high cost accuracy, low cost function, and has been trained with a high regularization value to avoid overfitting. I’m sure other data scientists have thought of the same idea, but if one could train a network with enough images of everything, and the network was exceptionally deep and complex, would it be able to characterize everything?

Another question I had while reading this is if the researchers had added a few more output nodes to the output layer, representing “nothing”, “noise”, “unrecognizable”, etc., would it have made a difference? Would allowing a network to choose “noise” a responsible choice since it might end up placing all unknown or confusing pictures in this instead of making a better guess. It’s like when working with kids, its better to have them choose something instead of nothing because they then learn to grow and develop themselves. If you present a kid with an escape option instead of a real substantial choice, it might nullify the results.

I found the fake images to have a vague and abstract similarity to the item the network was falsely identifying. For example, the network claimed that a penguin was come round black circles with a gray/white underneath. The network claimed that a picture of white and red stripes, similar to the colors of baseball, were indeed a baseball. I think it is unfair for the researchers to input the noisy images (looks like static on the television) and display the results. This is obviously going to produce an answer through the network. They never trained the network to have an option of “static noise” or “nothing” and therefore the network will give an arbitrary answer. This is because, at the end of the day, the inputs are just numbers, the outputs are just numbers and the inputs go through a series of matrix operations and are spit out on the other side. There is not meaning to the network because it is just completing a math operation. It is the job of the scientist and researcher to be responsible in how they train their networks and the data the network will ultimately have to categorize.