Analysis: Deep Neural Networks are Easily Fooled: High Confidence Predictions for Unrecognizable Images

This paper discusses the ease of generation of examples that can fool deep neural networks with high confidence. Basically, the paper shows that it is easy to produce images that are completely unrecognizable to humans, but state-of-the art DNNs believe to be recognizable objects with 99.99% confidence for example the guitar and penguin images as shown in the paper. The results provided in this paper are very important as random images pose a very successful adversarial attack especially if it is unrecognizable by human experts.

Experiment:

1. They chose Alex Net DNN which is trained on 1.2 million image ILSVRC 2012 ImageNet dataset- ImageNet DNN. LeNet model trained on MNIST dataset– MNIST DNN.
2. The modified images are generated by Evolutionary Algorithms, he images with highest probability will be selected and will undergo changes.
3. The EA's are tested with direct and indirect encodings, specifically, the indirect encoding here is a compositional pattern-producing network, which can evolve complex, regular images that resemble natural and man-made objects.
4. Fooling images via Gradient Ascent: found that images can be made that are also classified by DNNs with 99.99% confidence, despite them being mostly unrecognizable.

Results:

1. ImageNet: with overrepresented negative class the median confidence score decreased from 88.1% for DNN1 to 11.7% for DNN2.
2. MNIST: evolution produced many unrecognizable images with high confidence score even after 15 iterations, negative class is overrepresented (25% of the training set).

Conclusion:

Two different ways of encoding evolutionary algorithms produce two qualitatively different types of unrecognizable "fooling images", and gradient ascent produces a third.

Likes:

1. The paper provides an evolutionary model to generate high confidence examples for both ImageNet and MNIST datasets and this paper shows examples that fool one DNN are capable of fooling others.
2. The paper also highlights the areas of concerns like a security camera that relies on face or voice recognition being compromised by swapping white noise for a face, fingerprints. The fact that DNNs are increasingly used in a wide variety of industries, including safety-critical ones such as driverless cars, raises the possibility of costly exploits via techniques that generate fooling images.
3. What I also liked about this paper was their statement about the differences in the way Humans and DNNs recognise images and objects. There is always a comparison between humans and DNNs and they described that these DNNs can be fooled.

Dislikes:

1. In this paper they tell that it is not easy to prevent the DNNs from being fooled by retraining them with fooling images. I think this was the point where the future work prospects come in. I would have explored more to find on how to prevent these DNNs from being fooled.
2. The discussion section also shows that they try to propose an explanation about their results. I always feel that as a researcher one should back their results with good theory.
3. I agree that their results are impressive, and they highlight a very important safety issue when using DNNs but somewhere I also think that they overgeneralize DNNs. They give examples and a model to fool DNNs and they try to apply this to all the DNNs.

I have started exploring more and reading more papers now. It takes me significant time to understand a paper and after reading a paper I am mostly impressed by the results. These papers are always inspiring to me. As a researcher their results look good to place trust in them.