Deep Neural Networks Are Easily Fooled: High Confidence Predictions for Unrecognizable Images - Review and Discussion

Rajat Vikram Singh
rajats@andrew.cmu.edu

I. Introduction

Deep neural networks' (DNNs) high accuracy and classification prowess makes them an attractive tool in any machine learning enthusiast's toolset. Some of the fields where DNNs are used heavily are computer vision, speech recognition, NLP etc. Different architectures and types of deep networks are used for solving different kinds of problems. For solving computer vision problems, a very popular kind of deep networks called convolutional neural networks are used. CNNs are very useful in classification objects (ImageNet), handwritten digits (MNIST), with the state of the art reaching 4% of errors in classification. DNNs work by carving the loss surface for each class in a high-dimensional space. In this paper the authors have showed that these deep networks can be fooled by getting some high confidence predictions for images which lie within these boundaries but are unrecognizable. The authors have made some interesting observations. Some of them are listed below in my own words. These observations are followed by a discussion of how they fit in with the intuition of deep networks.

II. Interesting Observations

Following are some of the interesting observations from the paper:

a. Working with indirectly coded images, we see that there are some patterns which are recognizable. As the authors say, these images are recognizable if the original class of the data is known. The authors explain this observation by saying that these images contain the descriptive rare features which make it easy for the CNN to recognize.
b. An important result is that the directly encoded images were not able to get good accuracy in the case of ImageNet dataset (compared to the MNIST database) which contains a considerable amount of more samples, which can allude to the fact that having well-represented dataset can be a little difficult to fool.
c. Another important observation made by the authors was that DNNs are learning low and middle-level features rather than the global structure of the data which is showed when the accuracy of the DNN dropped when the authors removed some of the repeated elements of the image.
d. Using gradient ascent, the authors get similar results.

III. Discussion

The authors use evolutionary techniques to generate the fooling images. In the context of this paper, evolutionary algorithms work by keeping the best features of the image and perturbing the rest. The best features are selected using a fitness function. The fitness function used in this paper's experiments is the highest prediction score generated by the DNN. The authors have used two different encoding methods: 1) direct encoded images which encodes pixel level information which is quite detailed and is the source of the salt and pepper noise which is visible in the fooling images, 2) indirect encoded images which are generated by another DNN and the best images are selected by humans, which causes the images to be much more recognizable than the directly encoded images. The authors also used gradient ascent in their experimentations. The basic intuition behind all these approaches is gradient ascent or moving away from the local minima and hypothesizing that images a little further away from the minima will still lie within the loss surface of a particular class and will look like the images in that class. Evolutionary algorithms keep the rare differentiating feature in the image and perturb others, moving further from the class minima and moving away from it in a manner which makes intuitive sense. In a way, this approach could be thought of reverse dropout where the features with the biggest effect are kept in an attempt to over-fit the data. The following image from the paper explains the experimentation:

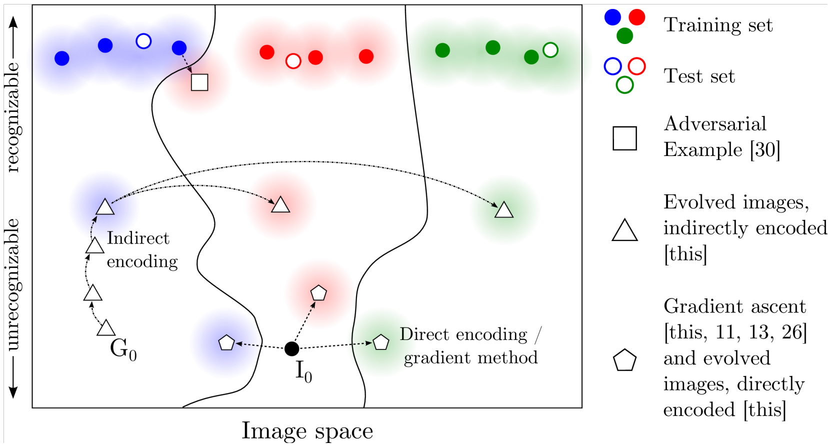
recognizable
unrecognizable
Indirect encoding
G_0
I_0
Direct encoding / gradient method
Image space
Training set
Test set
Adversarial Example [30]
Evolved images, indirectly encoded [this]
Gradient ascent [this, 11, 13, 26] and evolved images, directly encoded [this]

Fig. 1: Decision boundaries and generated examples

As shown in the legend, the bold blue, green and red circles are the initial training samples and the solid lines are the learned decision boundary by the DNN. We move away from the training set using evolutionary and gradient ascent methods generating the triangle and pentagon data points. However, these instances although lying within the decision boundaries of a particular class lie far away from the group of recognizable images (training set, solid circles). These images are the fooling images and confirm the authors' hypothesis that DNNs can be fooled by using these fooling images. There are multiple reasons for this kind of behavior:

a. This behavior hinges on the fact that if some data belongs to a class statistically, it does not mean that the pixel data makes human sense as well. It is interesting that these images pass through CNNs without containing any visual features of their classes since we are using convolutional neural networks which focus on image features.
b. Evolving directly encoded images produce white-noise static images. There might be some set of pixels which can be seen as creating features which the CNN is looking for but is not at all apparent to the human comprehension.
c. Given the nature of the evolutionary algorithms, only the discriminative features are retained between evolutions, making some of the features visible in the case of irregularly encoded images.

The authors also explain the well-cited paper by Szegedy et al [2] which showed that CNNs misclassify an image from the training set if imperceptible modifications are made to the image. This is an important result since it questions the regularization of the DNNs and questions the usefulness of the high accuracy these CNNs get on datasets like ImageNet. These examples are called adversarial examples. Authors show these adversarial examples in the figure above with a square data-point which lies in the class boundaries of a different class but is close to the images from it's original class since it is recognizable. A lot of research has ensued after the discovery of these adversarial examples.

IV. Conclusion

The results of the adversarial examples paper by this paper are relevant and we need to study this problem more. There are many scenarios where DNNs can be fooled by adversaries to get a specific output from the DNN.

As part of researching this problem, I came across some other interesting papers which talk about similar fooling of DNNs. These papers are added in the references section. It is an interesting and active research topic with big deep learning names associated with it.

V. References

[1] A. Nguyen, J. Yosinski, J. Clune. Deep Neural Networks Are Easily Fooled: High Confidence Predictions for Unrecognizable Images.

[2] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus. Intriguing properties of neural networks.

[3] I. J. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples.

[4] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio. Generative Adversarial Nets.