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

Xuewen Yang

August 8 2018

Abstract

Deep neural networks (DNNs) have recently been achieving state-of-the-art performance on a variety of pattern-recognition tasks, most notably visual classification problems. Given that DNNs are now able to classify objects in images with near-human-level performance, questions naturally arise as to what differences remain between computer and human vision. A recent study revealed that changing an image (e.g. of a lion) in a way imperceptible to humans can cause a DNN to label the image as something else entirely (e.g. mislabeling a lion a library). Here the author show a related result: it is easy to produce images that are completely unrecognizable to humans, but that state-of-theart DNNs believe to be recognizable objects with 99.99% confidence (e.g. labeling with certainty that white noise static is a lion). Specifically, they take convolutional neural networks trained to perform well on either the ImageNet or MNIST datasets and then find images with evolutionary algorithms or gradient ascent that DNNs label with high confidence as belonging to each dataset class. It is possible to produce images totally unrecognizable to human eyes that DNNs believe with near certainty are familiar objects, which they call fooling images (more generally, fooling examples). Their results shed light on interesting differences between human vision and current DNNs, and raise questions about the generality of DNN computer vision.

1. Introduction

Deep neural networks (DNNs) learn hierarchical layers of representation from sensory input in order to perform pattern recognition [1]. Recently, these deep architectures have demonstrated impressive, state-of-the-art, and sometimes human-competitive results on many pattern recognition tasks, especially vision classification problems [3]. Given the near-human ability of DNNs to classify visual objects, questions arise as to what differences remain between computer and human vision.

A recent study revealed a major difference between DNN and human vision. Changing an image, originally correctly classified (e.g. as a lion), in a way imperceptible to human eyes, can cause a DNN to label the image as something else entirely (e.g. mislabeling a lion a library).

2. Methods

2.1. Deep neural network models

To test whether DNNs might give false positives for unrecognizable images, they need a DNN trained to near state-of-the-art performance. They choose the well-known AlexNet architecture from [2], which is a convnet trained on the 1.3-million-image ILSVRC 2012 ImageNet dataset. Specifically, they use the already-trained AlexNet DNN provided by the Caffe software package. It obtains 42.6% top-1 error rate, similar to the 40.7% reported by Krizhevsky 2012 [2]. While the Caffe-provided DNN has some small differences from Krizhevsky 2012, they do not believe our results would be qualitatively changed by small architectural and optimization differences or their resulting small performance improvements. Similarly, while recent papers have improved upon Krizhevsky 2012, those differences are unlikely to change our results. We chose AlexNet because it is widely known and a trained DNN similar to it is publicly available. In this paper, they refer to this model as ImageNet DNN.

2.2. Generating images with evolution

They test EAs with two different encodings, meaning how an image is represented as a genome. The first has a direct encoding, which has one grayscale integer for each of 28×28 pixels for MNIST, and three integers (H, S, V) for each of 256×256 pixels for ImageNet. Each pixel value is initialized with uniform random noise within the [0; 255] range. Those numbers are independently mutated; first by determining which numbers are mutated, via a rate that starts at 0.1 (each number has a 10% chance of being chosen to be mutated) and drops by half every 1000 generations. The numbers chosen to be mutated are then altered via the polynomial mutation operator with a fixed mutation strength of 15. The second EA has an indirect encoding, which is more likely to produce regular images, meaning images that contain compressible patterns (e.g. symmetry and repetition). Indirectly encoded images tend to be regular because elements in the genome can affect multiple parts of the image. Specifically, the indirect encoding here is a compositional pattern-producing network (CPPN), which can evolve complex, regular images that resemble natural and man-made objects.

References

[1] G. E. Hinton. Learning multiple layers of representation. *Trends in Cognitive Sciences*, 2007. 1

[2] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012. 1

[3] Q. V. Le, W. Y. Zou, S. Y. Yeung, and A. Y. Ng. Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis. In *CVPR*, 2011. 1