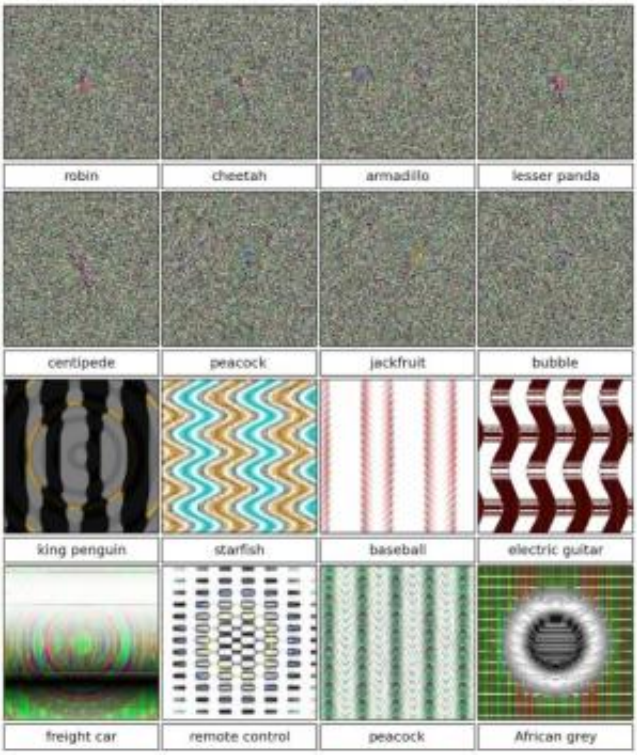
Researchers find a way to fool deep neural networks into 'recognizing' images that aren't there

12 December 2014, by Bob Yirka



Evolved images that are unrecognizable to humans, but that state-of-the-art DNNs trained on ImageNet believe with ? 99.6% certainty to be a familiar object. This result highlights differences between how DNNs and humans recognize objects. Images are either directly (top) or indirectly (bottom) encoded. Credit: *arXiv*:1412.1897

A trio of researchers in the U.S. has found that deep neural networks (DNNs) can be tricked into "believing" an image it is analyzing is of something recognizable to humans when in fact it isn't. They have written a paper about what they have discovered and uploaded it to the preprint server *arXiv*.

As time marches on, we humans are becoming more accustomed to computers being able to recognize things around us (faces on our smartphones, for example) and to do something with that information (pick out the face of a wanted person from a crowd). As part of that process we've come to believe that such systems are as good as they seem. But, as the trio working on this new effort has found, that assessment may be incorrect.

DNNs "learn" to recognize images by being exposed to many of those of the same type (thousands or millions of faces, for example)—they use learning algorithms that spot commonalities between parts of information in the images to map out different aspects of different objects. Once the learning has progressed to a certain level, the DNN is able to very accurately predict what object appears in an image, except, apparently, under certain circumstances. To find this rare circumstance, the researchers hooked a well known and respected DNN called AlexNet to a system that also included algorithms developed to evolve pictures using basic elements. The team expected the output to be exceptionally clear images of objects that most any person would instantly recognize. Instead, in many case, the result was a garbled mess, which the researchers described as static. More interesting, AlexNet offered confidence ratings up to 99 percent of the false images—the computer was certain the images were of things like lions, yet to the humans, they looked like static on an old TV set.

The reason for this apparent weakness in the DNN goes back to the way that they learn—all those parts that are supposed to add up to a discernible whole. If the algorithms creating the images add all the basic parts the DNN is looking for, but not in a logical way, then the result can look like static to people looking at them while appearing to be what the DNN learned from its early training, because it's able to find those basic image parts.

This discovery is more than just passing interest, DNNs are used in applications such as by cars that drive themselves—if someone with ill intent were bent on harm, it might not be too difficult to imagine placing something on a highway that to us humans looked like fog or smoke, but to the car's computer, was identified as an object or perhaps a pedestrian, causing an accident to occur as the car sought to prevent a collision with the nonexistent object.

More information: Deep Neural Networks are Easily Fooled: High Confidence Predictions for Unrecognizable Images, *arXiv*:1412.1897 [cs.CV] arxiv.org/abs/1412.1897

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 we show a related result: it is easy to produce images that are completely unrecognizable to humans, but that state-of-the-art DNNs believe to be recognizable objects with 99.99% confidence (e.g. labeling with certainty that white noise static is a lion). Specifically, we 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. Our results shed light on interesting differences between human vision and current DNNs, and raise questions about the generality of DNN computer vision.

neural-networks-images.html