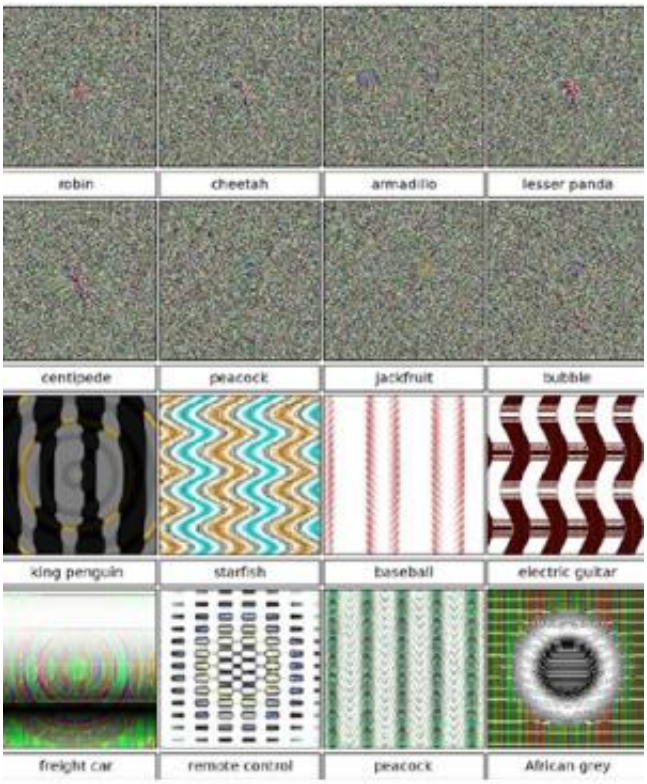
Images that fool computer vision raise security concerns

March 23 2015, by Bill Steele



Meaningless to humans, these images are recognized by a computer wiith great certainty as common objects. The white noise versions on top and pattern versions below were created by slightly different methods.

Computers are learning to recognize objects with near-human ability. But Cornell researchers have found that computers, like humans, can be fooled by optical illusions, which raises security concerns and opens new

avenues for research in computer vision.

Cornell graduate student Jason Yosinski and colleagues at the University of Wyoming Evolving Artificial Intelligence Laboratory have created images that look to humans like white noise or random geometric patterns but which computers identify with great confidence as common objects. They will report their work at the IEEE Computer Vision and Pattern Recognition conference in Boston June 7-12.

"We think our results are important for two reasons," said Yosinski. "First, they highlight the extent to which computer vision systems based on modern supervised machine learning may be fooled, which has security implications in many areas. Second, the methods used in the paper provide an important debugging tool to discover exactly which artifacts the networks are learning."

Computers can be trained to recognize images by showing them photos of objects along with the name of the object. From many different views of the same object the computer assembles a sort of fuzzy model that fits them all and will match a new image of the same object. In recent years, computer scientists have reached a high level of success in image recognition using systems called Deep Neural Networks (DNN) that simulate the synapses in a human brain by increasing the value of a location in memory each time it is activated. "Deep" networks use several layers of simulated neurons to work at several levels of abstraction: One level recognizes that a picture is of a four-legged animal, another that it's a cat, and another narrows it to "Siamese."

But computers don't process images the way humans do, Yosinski said. "We realized that the neural nets did not encode knowledge necessary to produce an image of a fire truck, only the knowledge necessary to tell fire trucks apart from other classes," he explained. Blobs of color and patterns of lines might be enough. For example, the computer might say

"school bus" given just yellow and black stripes, or "computer keyboard" for a repeating array of roughly square shapes.

Working in the Cornell Creative Machines lab with Hod Lipson, associate professor of mechanical and aerospace engineering, the researchers "evolved" images with the features a DNN would consider significant. They tested with two widely used DNN systems that have been trained on massive image databases. Starting with a random image, they slowly mutated the images, showing each new version to a DNN. If a new image was identified as a particular class with more certainty than the original, the researchers would discard the old version and continue to mutate the new one. Eventually this produced images that were recognized by the DNN with over 99 percent confidence but were not recognizable to human vision.

"The research shows that it is possible to 'fool' a deep learning system so it learns something that is not true but that you want it to learn," said Fred Schneider, the Samuel B. Eckert Professor of Computer Science and a nationally recognized expert on computer security. "This potentially has the basis for malfeasants to cause automated systems to give carefully crafted wrong answers to certain questions. Many systems on the Web are using deep learning to analyze and draw inferences from large sets of data. DNN might be used by a Web advertiser to decide what ad to show you on Facebook or by an intelligence agency to decide if a particular activity is suspicious."

Malicious Web pages might include fake images to fool image search engines or bypass "safe search" filters, Yosinski noted. Or an apparently abstract image might be accepted by a facial recognition system as an authorized visitor.

In a further step, the researchers tried "retraining" the DNN by showing it fooling images and labeling them as such. This produced some

improvement, but the researchers said that even these new, retrained networks often could be fooled.

"The field of image recognition has been revolutionized in the last few years," Yosinski said. "[Machine learning researchers] now have a lot of stuff that works, but what we don't have, what we still need, is a better understanding of what's really going on inside these neural networks."

Provided by Cornell University