How to Fool a Computer With Optical Illusions

By Adrienne LaFrance

Computers, like people, understand what they see in the world based on what they've seen before.

And computer brains have become really, really good at being able to identify all kinds of things. Machines can recognize faces, read handwriting, interpret EKGs, even describe what's happening in a photograph. But that doesn't mean that computers *see* all those things the same way that people do.

This might sound like a throwaway distinction. If everybody—computers and humans alike—can see an image of a lion and call it a lion, what does it matter how that lion looks to the person or computer processing it? And it's true that ending up at the same place can be more useful than tracing how you got there. But to a hacker hoping to exploit an automated system, understanding an artificial brain's way of seeing could be a way in.

A team of computer scientists from the University of Wyoming and Cornell University recently figured out how to create a whole class of images that appear meaningful to computers but look like TV static or glitch art to the human eye. "It is easy to produce images that are completely unrecognizable to humans, but that state-of-the-art [deep neural networks] believe to be recognizable objects," they wrote in a paper that's currently under peer review and has been posted to ArXiv, where scientists post preprints of papers while they are being reviewed.

And not only do computers recognize signals in the noise, they do so with a huge amount of confidence. So that while you see images that look like this...

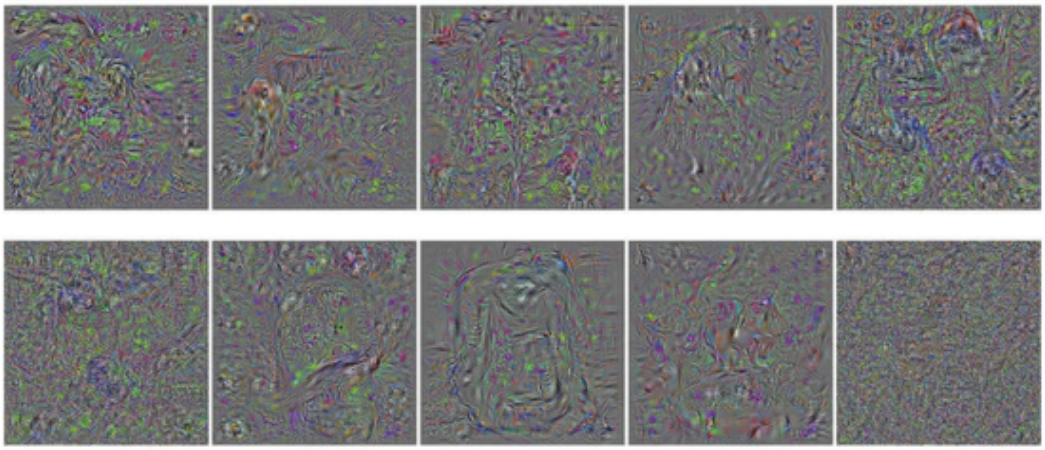
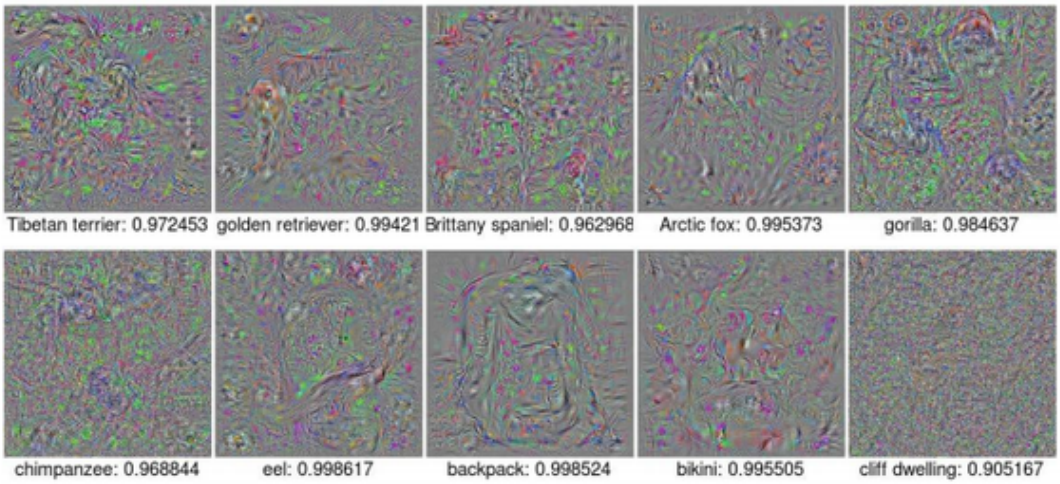


Image derived from Nguyen, Yosinski, Clune paper

...a computer's brain, or deep neural network (DNN), says it is 99 percent sure that it sees in those same images a gorilla, and an arctic fox, and a bikini, and an eel, and a backpack, and so on.



Screenshot from Nguyen, Yosinski, Clune paper

"To some extent these are optical illusions for artificial intelligence," co-author Jeff Clune told me via gchat. "Just as optical illusions exploit the particular way humans see... these images reveal aspects of how the DNNs see that [make] them vulnerable to being fooled, too. But DNN optical illusions don't fool us because our vision system is different."

Clune and his team used an algorithm to generate random images that appeared unrecognizable to humans. At first, Clune explains, the computer might be unsure about what it was seeing: "It then says, 'That doesn't look like much of anything, but if you forced me to guess, the best I see there is a lion. But it only 1 percent looks like a lion.'"

From there, the researchers would continue to randomly tweak the image's pixels—which remained unidentifiable to humans—until the computer said it could identify, with almost complete certainty, the image as a familiar object. And though the image would still appear nonsensical to the human eye, it would represent, Clune says, the Platonic form of whatever the computer sees. And this is a key point: Because it's not that the computer is identifying the image incorrectly per se, it's that a computer sees and thinks about the identifying components of any given thing differently—and more granularly—than a human does. "One way to think about it is this," Clune told me. "These DNNs are fans of cubist art. They want to see an eye, a nose, and a mouth in the image to call it a face, but they don't particularly care where those things are. The mouth can be above the eyes and to the left of the nose."

But while humans squinting at the same blocks of color in a Paul Klee painting might identify different familiar objects—what looks like a duck to me might look like a rabbit to you—DNNs will look at the same seemingly abstract image and derive the same meaning. "We tried exactly that and it works," Clune said of testing the same illusion against multiple neural networks. "Two different DNNs will both look at the same TV-static and say, 'Yep. Definitely a lion.'"

A human looking at a lion is making split-second categorizations as electrical signals travel along the optic nerve to the brain: *Okay, it's an animal. It's big. It walks on four legs. It has a tail. It has a sandy mane—oh, that's a lion*. A computer brain's checklist is more refined. Instead of looking for retractable claws and sharp teeth, artificial intelligence assesses *lionishness* at the pixel level. Which

means an image that looks like a snowy monitor to a human brain may look quite clearly like a big cat to a computer brain, sort of like how you might see (or not) a hidden sailboat in a Magic Eye poster.

And because these computers see illusions the same way, there are significant implications for digital security, surveillance, even human communications. "For example, Google image search filters out X-rated images automatically," Clune said. "Using the technique we describe, a shady company could make images that look to Google's artificial intelligence filters like rabbits, but that actually contain nudity or other illicit imagery."

To people in countries where governments restrict speech and publishing, citizens could theoretically communicate secretly by leveraging the opacity within deep neural networks. Clune: "People could embed messages discussing freedom of the press and get them past communist AI-censoring filters by making the image look like the communist party flag!"

Even when computers can be trained that what they're seeing isn't, as far as a human is concerned, the thing the computer thinks it sees—it's easy to retrain the computer to be fooled all over again, which, for now, leaves such networks vulnerable to hackers. Understanding such opportunities for exploitation will be critical as artificial intelligence becomes increasingly pervasive.

In the meantime, Clune says his team's findings have underscored limitations in the human way of seeing. "This work has caused me to reflect on how we see even more deeply," he said. "Do I focus on the low-level details only sometimes? Only on the high-level structure and ignore the details?"

Exercises in perspective aside, the larger promise of deep neural networks, Clune says, is astonishing. "They have already, and will—more than you can imagine—change the course of human history."

This article available online at:

http://www.theatlantic.com/technology/archive/2014/12/how-to-fool-a-computer-with-optical-illusions/383779/