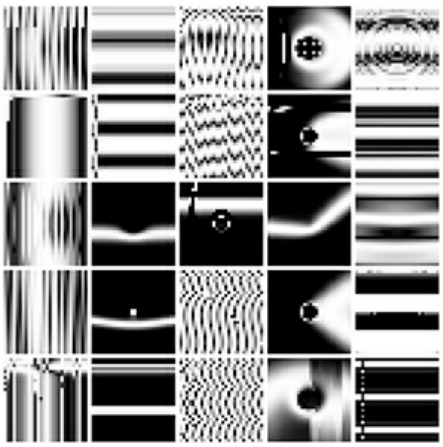
COMPUTING NEWS

8 COMMENTS

"Smart" Software Can Be Tricked into Seeing What Isn't There

Humans and software see some images differently, pointing out shortcomings of recent breakthroughs in machine learning.

By Caleb Garling on December 24, 2014



Images like these were created to trick machine learning algorithms. The software sees each pattern as one of the digits 1 to 5.

A technique called deep learning has enabled Google and other companies to make breakthroughs in getting computers to understand the content of photos. Now researchers at Cornell University and the University of Wyoming have shown how to make images that fool such software into seeing things that aren't there.

The researchers can create images that appear to a human as scrambled nonsense or simple geometric patterns, but are identified by the software as an everyday object such as a school bus. The trick images offer new insight into the differences between how real brains and the simple simulated neurons used in deep learning process images.

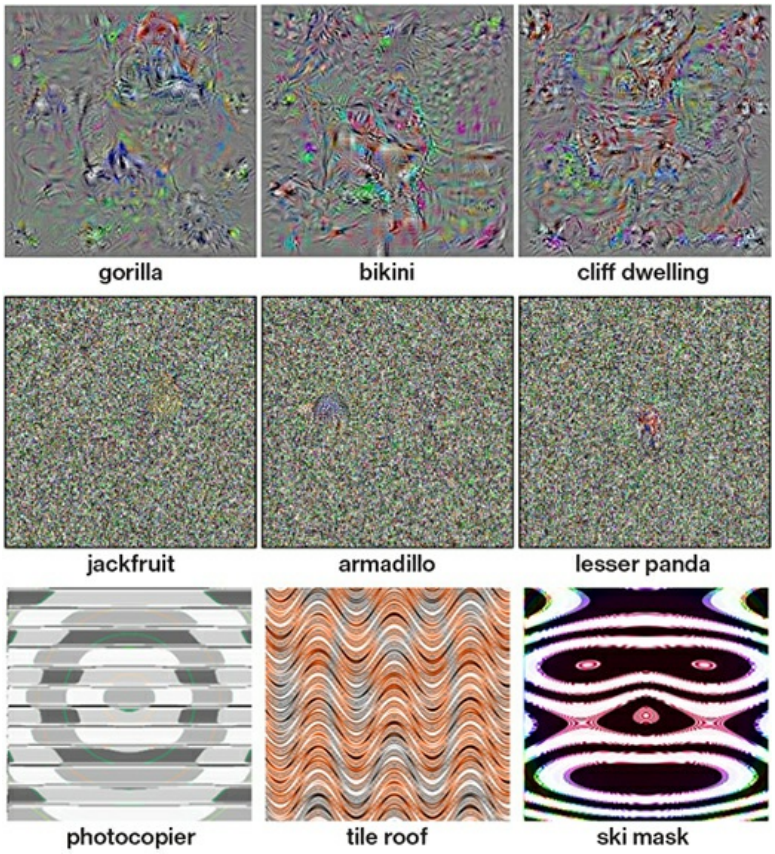
Researchers typically train deep learning software to recognize something of interest—say, a guitar—by showing it millions of pictures of guitars, each time telling the computer "This is a guitar." After a while, the software can identify guitars in images it has never seen before, assigning its answer a confidence rating. It might give a guitar displayed alone on a white background a high confidence rating, and a guitar seen in the background of a grainy cluttered picture a lower confidence rating (see "10 Breakthrough Technologies 2013: Deep Learning").

That approach has valuable applications such as facial recognition, or using software to process security or traffic camera footage, for example to measure traffic flows or spot suspicious activity.

WHY IT MATTERS

Image recognition algorithms are becoming widely used in many products and services.

But although the mathematical functions used to create an artificial neural network are understood individually, how they work together to decipher images is unknown. “We understand that they work, just not how they work,” says Jeff Clune, an assistant professor of computer science at the University of Wyoming. “They can learn to do things that we can’t even learn to do ourselves.”



These images look abstract to humans, but are seen by the image recognition algorithm they were designed to fool as the objects described in the labels.

To shed new light on how these networks operate, Clune’s group used a neural network called AlexNet that has achieved impressive results in image recognition. They operated it in reverse, asking a version of the software with no knowledge of guitars to *create* a picture of one, by generating random pixels across an image.

The researchers asked a second version of the network that had been trained to spot guitars to rate the images made by the first network. That confidence rating was used by the first network to refine its next attempt to create a guitar image. After thousands of rounds of this between the two pieces of software, the first network could make an image that the second network recognized as a guitar with 99 percent confidence.

However, to a human, those “guitar” images looked like colored TV static or simple patterns. Clune says this shows that the software is not interested in piecing together structural details like strings or a fretboard, as a human trying to identify something might be. Instead, the software seems to be looking at specific distance or color relationships between pixels, or overall color and texture.

That offers new insight into how artificial neural networks really work, says Clune, although more research is needed.

Ryan Adams, an assistant computer science professor at Harvard, says the results aren’t completely surprising. The fact that large areas of the trick images look like seas of static probably stems from the way networks are fed training images. The object of interest is usually only a small part of the photo, and the rest is unimportant.

Adams also points out that Clune’s research shows

humans and artificial neural networks do have some things in common. Humans have been thinking they see everyday objects in random patterns—such as the stars—for millennia.

Clune says it would be possible to use his technique to fool image recognition algorithms when they are put to work in Web services and other products. However, it would be very difficult to pull off. For instance, Google has algorithms that filter out pornography from the results of its image search service. But to create images that would trick it, a prankster would need to know significant details about how Google's software was designed.

Credit: Images courtesy of University of Wyoming

Tagged: Computing

RELATED STORIES YOU MAY HAVE MISSED MORE FROM THIS AUTHOR

Auras: There's an App for That

Hundreds of Portuguese Buses and Taxis Are Also Wi-Fi Routers

Singapore Wants a Driverless Version of Uber

2015 Could Be the Year of the Hospital Hack

12

The Startup Meant to Reinvent What Bitcoin Can Do

21

AT&T Builds an Assistant App with Social Skills

Researchers Will Study Police Confrontations Via Body Cameras

HP Will Release a "Revolutionary" New Operating System in 2015

53

Feeling More Amateur Than Maestro with a Finger-Worn Mouse

THE LATEST POPULAR MOST SHARED