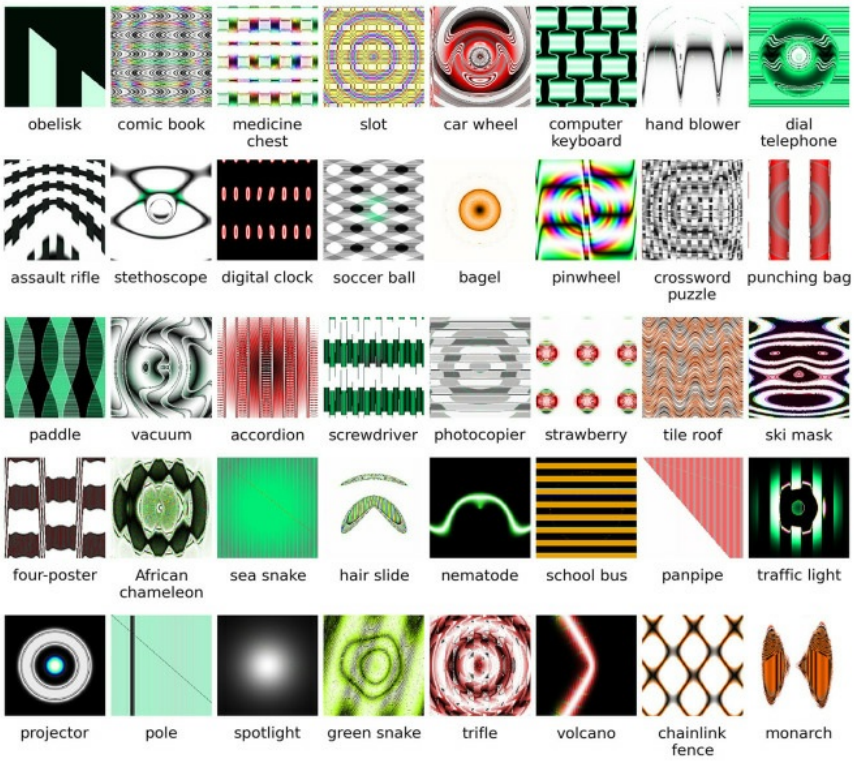
DESIGN

Simple Pictures That State-of-the-Art AI Still Can't Recognize

BY KYLE VANHEMERT 01.05.15 | 6:30 AM | PERMALINK



Courtesy Jeff Clune

Look at these black and yellow bars and tell me what you see. Not much, right? Ask state-of-the-art artificial intelligence the same question, however, and it will tell you they're a school bus. It will be over 99 percent certain of this assessment. And it will be totally wrong.

Computers are getting truly, freakishly good at identifying what they're looking at. They can't look at this picture and tell you it's a chihuahua wearing a sombrero, but they can say that it's a dog wearing a hat with a wide brim. A new paper, however, directs our attention to one place these super-smart algorithms are totally stupid. It details how researchers were able to fool cutting-edge deep neural networks using simple, randomly generated imagery. Over and over, the algorithms looked at abstract jumbles of shapes and thought they were seeing parrots, ping pong paddles, bagels, and butterflies.

The findings force us to acknowledge a somewhat obvious but hugely

MOST RECENT WIRED POSTS


Watch Code Warp Peoples' Hands Into Trippy Visuals


Sennheiser Adds Wireless and Noise-Canceling Features to Its Consumer Headphones


Razer's New Virtual Reality Gaming Headset Encourages Open-Source Hacking


Rap Weirdos on This Week's Playlist Start 2015 on the Right Note


Razer Debuts Android Set-Top Box That Also Streams PC Games to Your TV


Turns Out the Internet Is Bad at Guessing How Many Coins Are in a Jar

TRENDING NOW ON WIRED

Bill Gates' Plan to Help the Developing World Profit From Its Sewage

11 More Cool Gadgets From CES: Robot Plant Feeders, Bossy Yoga Mats, and Alien Speakers

Our System Is So Broken, Almost No Patented Discoveries Ever Get Used

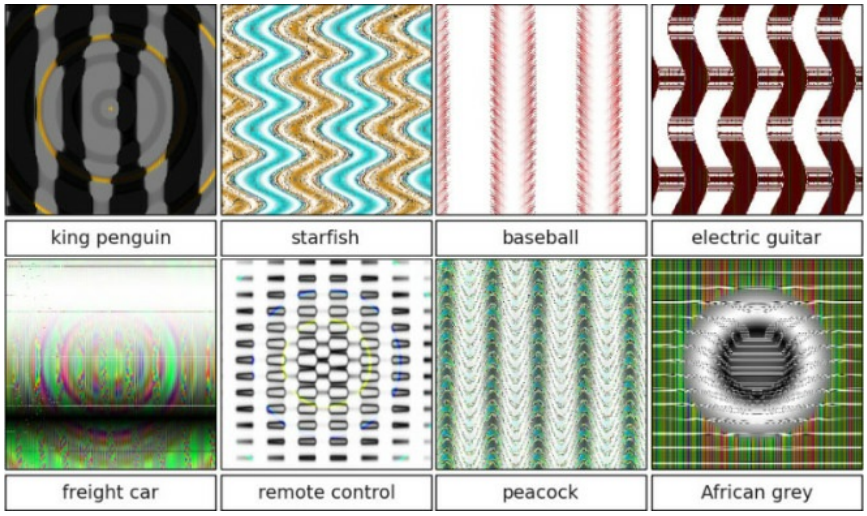
Never Buy a Phone Again

You Should Be Watching The Fall, a Serial-Killer Show Like No Other

important fact: Computer vision and human vision are nothing alike. And yet, since it increasingly relies on neural networks that teach themselves to see, we're not sure precisely how computer vision differs from our own. As Jeff Clune, one of the researchers who conducted the study, puts it, when it comes to AI, "we can get the results without knowing how we're getting those results."

Evolving Images to Fool AI

One way to find out how these self-trained algorithms get their smarts is to find places where they are dumb. In this case, Clune, along with PhD students Anh Nguyen and Jason Yosinski, set out to see if leading image-recognizing neural networks were susceptible to false positives. We know that a computer brain can recognize a koala bear. But could you get it to call something else a koala bear?



Nope. Nope. Nope. Nope. Nope. Nope. Nope. Nope. Courtesy Jeff Clune

To find out, the group generated random imagery using evolutionary algorithms. Essentially, they bred highly-effective visual bait. A program would produce an image, and then mutate it slightly. Both the copy and the original were shown to an "off the shelf" neural network trained on ImageNet, a data set of 1.3 million images, which has become a go-to resource for training computer vision AI. If the copy was recognized as something—anything—in the algorithm's repertoire with more certainty the original, the researchers would keep it, and repeat the process. Otherwise, they'd go back a step and try again. "Instead of survival of the fittest, it's survival of the prettiest," says Clune. Or, more accurately, survival of the most recognizable to a computer as an African Gray Parrot.

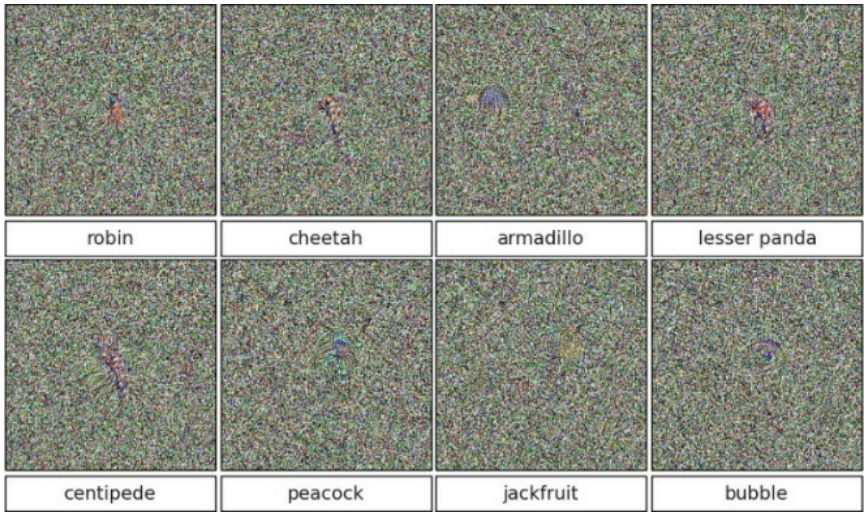
Eventually, this technique produced dozens images that were recognized by the neural network with over 99 percent confidence. To you, they won't seem like much. A series of wavy blue and orange lines. A mandala of ovals. Those alternating stripes of yellow and black. But to the AI, they were obvious matches: Star fish. Remote control. School bus.

Peering Inside the Black Box

In some cases, you can start to understand how the AI was fooled. Squint your eyes, and a school bus can look like alternating bands of yellow and

black. Similarly, you could see how the randomly generated image that triggered “monarch” would resemble butterfly wings, or how the one that was recognized as “ski mask” does look like an exaggerated human face.

But it gets more complicated. The researchers also found that the AI could routinely be fooled by images of pure static. Using a slightly different evolutionary technique, they generated another set of images. These all look exactly alike—which is to say, nothing at all, save maybe a broken TV set. And yet, state of the art neural networks pegged them, with upward of 99 percent certainty, as centipedes, cheetahs, and peacocks.



These static-like images also fooled the neural networks. Courtesy Jeff Clune

To Clune, the findings suggest that neural networks develop a variety of visual cues that help them identify objects. These cues might seem familiar to humans, as in the case of the school bus, or they might not. The results with the static-y images suggest that, at least sometimes, these cues can be very granular. Perhaps in training, the network notices that a string of “green pixel, green pixel, purple pixel, green pixel” is common among images of peacocks. When the images generated by Clune and his team happen on that same string, they trigger a “peacock” identification. The researchers were also able to elicit an identification of “lizard” with abstract images that looked nothing alike, suggesting that the networks come up with a handful of these cues for each object, any one of which can be enough to trigger a confident identification.

The fact that we’re cooking up elaborate schemes to trick these algorithms points to a broader truth about artificial intelligence today: Even when it works, we don’t always know how it works. “These models have become very big and very complicated and they’re learning on their own,” say Clune, who heads the Evolving Artificial Intelligence Laboratory at the University of Wyoming. “There’s millions of neurons and they’re all doing their own thing. And we don’t have a lot of understanding about how they’re accomplishing these amazing feats.”

Studies like these are attempts to reverse engineer those models. They aim to find the contours of the artificial mind. “Within the last year or two, we’ve started to really shine increasing amounts of light into this black box,” Clune explains. “It’s still very opaque in there, but we’re starting to get a glimpse of it.”

Why Does a Computer's Bad Eye Sight Matter, Anyway?

Earlier this month, Clune discussed these findings with fellow researchers at the Neural Information Processing Systems conference in Montreal. The event brought together some of the brightest thinkers working in artificial intelligence. The reactions sorted into two rough groups. One group—generally older, with more experience in the field—saw how the study made sense. They might've predicated a different outcome, but at the same time, they found the results perfectly understandable.

The second group, comprised of people who perhaps hadn't spent as much time thinking about what makes today's computer brains tick, were struck by the findings. At least initially, they were surprised these powerful algorithms could be so plainly wrong. Mind you, these were still people publishing papers on neural networks and hanging out at one of the year's brainiest AI gatherings.

To Clune, the bifurcated response was telling: It suggested a sort of generational shift in the field. A handful of years ago, the people working with AI were building AI. These days, the networks are good enough that researchers are simply taking what's out there and putting it to work. "In many cases you can take these algorithms off the shelf and have them help you with your problem," Clune says. "There is an absolute gold rush of people coming in and using them."

That's not necessarily a bad thing. But as more stuff is built on top of AI, it will only become more vital to probe it for shortcomings like these. If it really just takes a string of pixels to make an algorithm certain that a photo shows an innocuous furry animal, think how easy it could be to slip pornography undetected through safe search filters. In the short term, Clune hopes the study will spur other researchers to work on algorithms that take images' global structure into account. In other words, algorithms that make computer vision more like human vision.

But the study invites us to consider other forms these vulnerabilities could take. Does facial recognition, for instance, rely on the same sort of technology?

"Exactly the same," Clune says. "And it's susceptible to the exact same problem."

You can imagine all sorts of interesting implications here. Maybe a certain 3-D printed nose could enough to make a computer think you're someone else. Perhaps a mask of some precise geometry could render you invisible to a surveillance system entirely. A few years back, British design group ScanLAB Projects proposed a series of speculative objects that could subvert laser scanning of 3-D spaces, obscuring doorways or inventing phantom passageways. This new work just confirms that as the use of computer vision grows, the possibilities for subversion will follow.

More broadly, though, it's a reminder of a fast-emerging reality as we enter the age of self-learning systems. Today, we're still in control of the things we're building. But as they increasingly help build themselves, we shouldn't be surprised to find them complex to the point of opacity. "It's no longer lines of computer code written in a way a human would write them," Clune says. "It's almost like an economy of interacting parts, and the intelligence emerges out of that." We'll undoubtedly waste no time putting that intelligence to use. It's less clear how fully we'll understand it when we do.