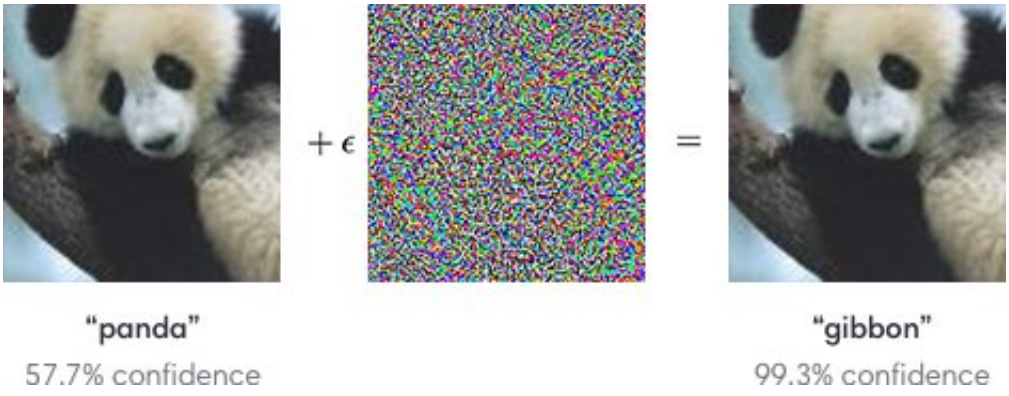
Fooling the human via changes to images

March 3 2018, by Nancy Owano



Credit: OpenAI

Well, so much for an assumption that now sounds too easy to accept—that the magnificent human brain has it over a machine any day. Really? Do we interpret the world more accurately than a "convolutional neural network" can?

As Even Ackerman pointed out, "when a CNN [convolutional neural network] is presented with an image, it's looking at a static grid of rectangular pixels."

We look at images and see them correctly, such as humans and animals; CNNs look at things more like computers.

A research team is raising questions about easy assumptions, however. They are exploring what happens with adversarial examples with regard

to humans.

Inputs to machine learning models designed to cause the models to make a mistake are "adversarial examples." Adversarial examples, as such, could potentially be dangerous.

Simply put, "Adversarial examples are malicious inputs designed to fool machine learning models," according to a Google Research page.

As a blog posting in OpenAI explained, attackers could target autonomous vehicles by using stickers or paint to create an adversarial stop sign that the vehicle would interpret as a 'yield' or other sign.

The researchers, in talking about machine learning models as vulnerable to adversarial examples, noted that small changes to images can cause computer vision models to make mistakes, such as identifying a school bus as an ostrich.

The blog from OpenAI referred to adversarial examples as representing a concrete problem in AI safety.

Having said that, what about adversarial examples fooling humans? Can that happen?

The team, said Even Ackerman in *IEEE Spectrum*, "decided to try and figure out whether the same techniques that fool artificial neural networks can also fool the biological neural networks inside of our heads."

The research paper describing their work is "Adversarial Examples that Fool both Human and Computer Vision," on arXiv.

"Here, we create the first adversarial examples designed to fool

humans," they wrote. They found that "adversarial examples that strongly transfer across computer vision models influence the classifications made by time-limited human observers." (Ackerman noted that in the study, people only had between 60 and 70 milliseconds to look at each image and make a decision.)

IEEE Spectrum's Even Ackerman discussed what they did and presented a set of two images from Google Brain to support his explanation.

Ackerman showed "a picture of a cat on the left. On the right, can you tell whether it's a picture of the same cat, or a picture of a similar looking dog? The difference between the two pictures is that the one on the right has been tweaked a bit by an algorithm to make it difficult for a type of computer model called a convolutional neural network (CNN) to be able to tell what it really is. In this case, the CNN thinks it's looking at a dog rather than a cat, but what's remarkable is that most people think the same thing."

What? How can humans make the same mistake? Ackerman said it might be possible to target the development of an adversarial image at humans "by choosing models that match the human visual system as closely as possible."

But what exactly is messing with the human's ability to be correct? Ackerman said the researchers pointed out that "our adversarial examples are designed to fool human perception, so we should be careful using subjective human perception to understand how they work."

He said they were willing to make some generalizations "about a few different categories of modifications, including 'disrupting object edges, especially by mid-frequency modulations perpendicular to the edge; enhancing edges both by increasing contrast and creating texture boundaries; modifying texture; and taking advantage of dark regions in

the image, where the perceptual magnitude of small perturbations can be larger.'"

How they tested: Subjects with normal or corrected vision participated in the experiment.

"For each group, a successful adversarial image was able to fool people into choosing the wrong member of the group, by identifying it as a dog when it's actually a cat, or vice versa," Ackerman said.

Subjects were asked to classify images that appeared on the screen by pressing buttons on a response time box, said the authors.

Ackerman wrote, "The short amount of time that the image was shown mitigated the difference between how CNNs perceive the world and how humans do."

The experiment involved three groups of images: pets (cats and dogs), vegetables (cabbages and broccoli), and "hazard" (spiders and snakes).

Ackerman's comment on the research findings was that "there's overlap between the perceptual manipulation of CNNs and the manipulation of humans. It means that machine learning techniques could potentially be used to subtly alter things like pictures or videos in a way that could change our perception of (and reaction to) them without us ever realizing what was going on."

He added that "we'll have to be careful, and keep in mind that just like those computers, sometimes we're far too easy to fool."

"Adversarial Examples that Fool both Human and Computer Vision" is by Gamaleldin F. Elsayed, Shreya Shankar, Brian Cheung, Nicolas Papernot, Alex Kurakin, Ian Goodfellow, and Jascha Sohl-Dickstein, on

arXiv.

More information: Adversarial Examples that Fool both Human and Computer Vision, arxiv.org/abs/1802.08195