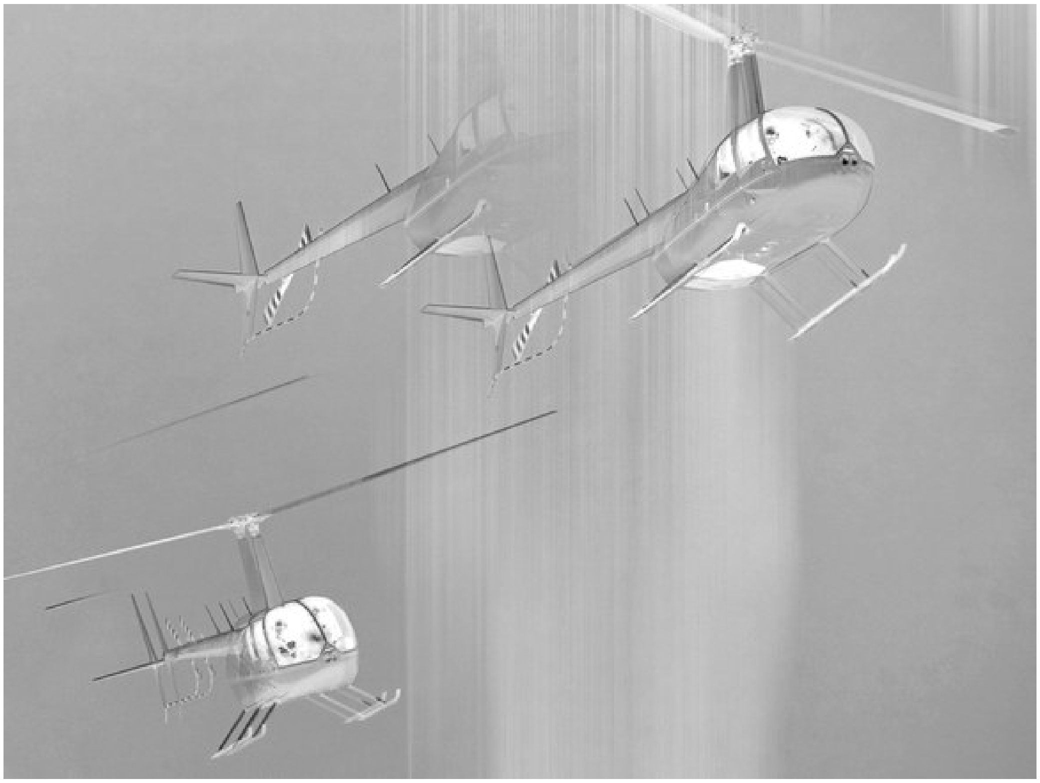
WIRED

LOUISE MATSAKIS SECURITY 12.20.17

RESEARCHERS FOOLED A GOOGLE AI INTO THINKING A RIFLE WAS A HELICOPTER



WIRED/GETTY IMAGES

TECH GIANTS LOVE to tout how good their computers are at identifying what's depicted in a photograph. In 2015, deep learning algorithms designed by Google, Microsoft, and China's Baidu superseded humans at the task, at least initially. This week, Facebook announced that its facial-recognition technology is now smart enough to identify a photo of you, even if you're not tagged in it.

But algorithms, unlike humans, are susceptible to a specific type of problem called an "adversarial example." These are specially designed optical illusions that fool computers into doing things like mistake a picture of a panda for one of a gibbon. They can be images, sounds, or paragraphs of text. Think of them as hallucinations for algorithms.

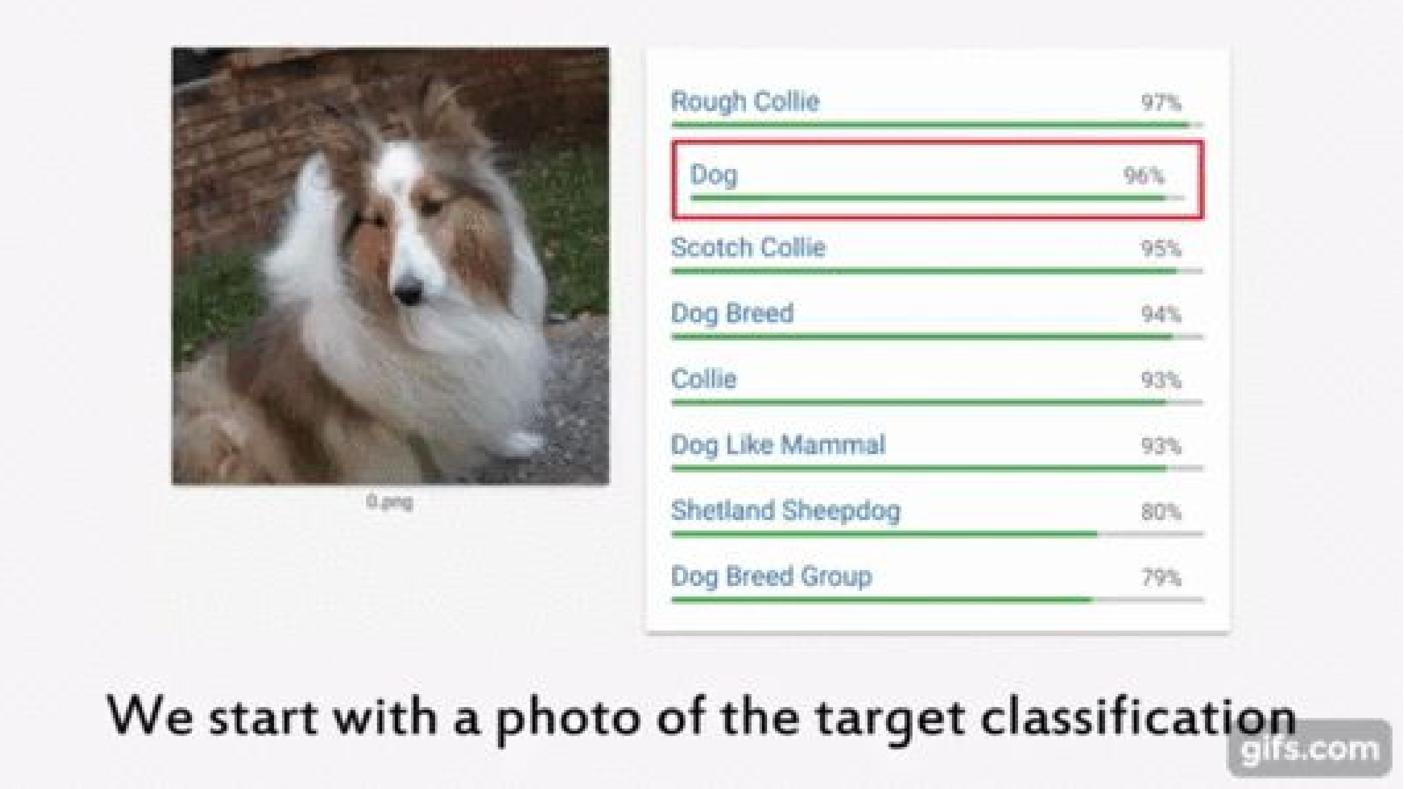
While a panda-gibbon mix-up may seem low stakes, an adversarial example could thwart the AI system that controls a self-driving car, for instance, causing it to mistake a stop sign for a speed limit one. They've already been used to beat other kinds of algorithms, like spam filters.

Those adversarial examples are also much easier to create than was previously understood, according to research released Wednesday from MIT's Computer Science and Artificial Intelligence Laboratory. And not just under controlled conditions; the team reliably fooled Google's Cloud Vision API, a machine learning algorithm used in the real world today.

Previous adversarial examples have largely been designed in "white box" settings, where computer scientists have access to the underlying mechanics that power an algorithm. In these scenarios, researchers learn how the computer system was trained, information that helps them figure out how to trick it. These kinds of adversarial examples are considered less threatening, because they don't closely resemble the real world, where an attacker wouldn't have access to a proprietary algorithm.

For example, in November another team at MIT (with many of the same researchers) published a study demonstrating how Google's InceptionV3 image classifier could be duped into thinking that a 3-D-printed turtle was a rifle. In fact, researchers could manipulate the AI into thinking the turtle was any object they wanted. While the study demonstrated that adversarial examples can be 3-D objects, it was conducted under white-box conditions. The researchers had access to how the image classifier worked.

But in this latest study, the MIT researchers did their work under "black box" conditions, without that level of insight into the target algorithm. They designed a way to quickly generate black-box adversarial examples that are capable of fooling different algorithms, including Google's Cloud Vision API. In Google's case, the MIT researchers targeted the part of the system of that assigns names to objects, like labeling a photo of a kitten "cat."



Rough Collie	97%
Dog	96%
Scotch Collie	95%
Dog Breed	94%
Collie	93%
Dog Like Mammal	93%
Shetland Sheepdog	80%
Dog Breed Group	79%

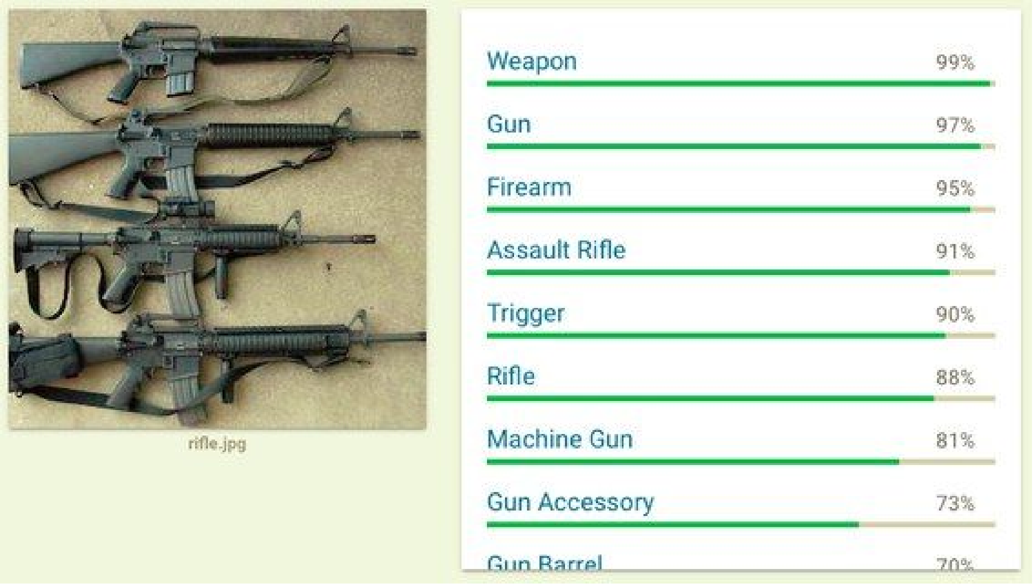
We start with a photo of the target classification

What it looks like when MIT's system attacks Google's algorithm. MIT

Despite the strict black box conditions, the researchers successfully tricked Google's algorithm. For example, they fooled it into believing a photo of a row of machine guns was instead a picture of a helicopter, merely by slightly tweaking the pixels in the photo. To the human eye, the two images look identical. The indiscernible difference only fools the machine.

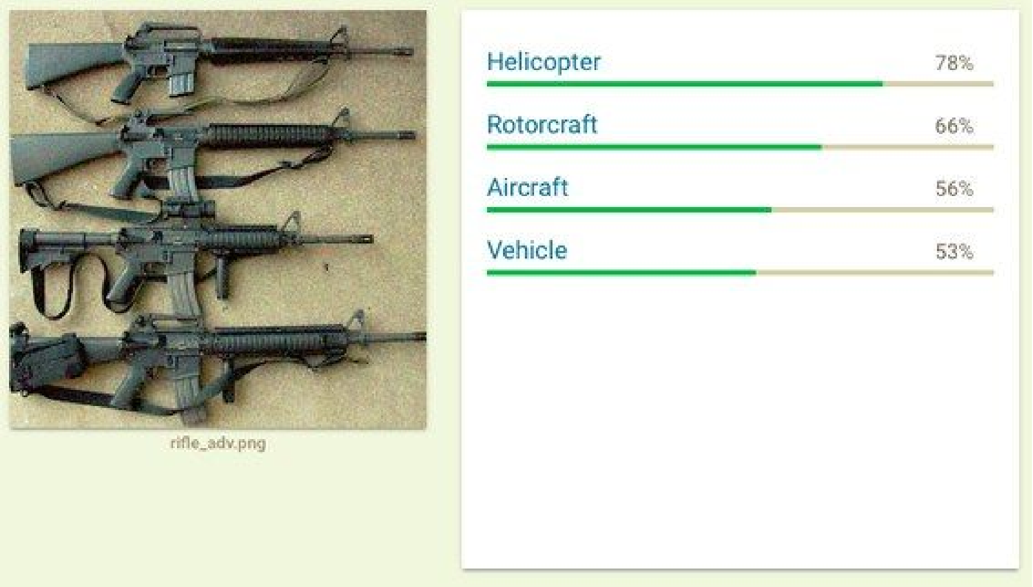
The researchers didn't just tweak the photos randomly. They targeted the AI system using a standard method. Each time they tried to fool the AI, they analyzed their results, and then intelligently inched toward an image that could trick a computer into thinking a gun (or any other object) is something it isn't.

The researchers randomly generated their labels; in the rifle example, the classifier "helicopter" could just as easily have been "antelope." They wanted to prove that their system worked, no matter what labels were chosen. "We can do this given anything. There's no bias, we didn't choose what was easy," says Anish Athalye, a PhD student at MIT and one of the lead authors of the paper. Google declined to comment in time for publication.



Weapon	99%
Gun	97%
Firearm	95%
Assault Rifle	91%
Trigger	90%
Rifle	88%
Machine Gun	81%
Gun Accessory	73%
Gun Barrel	70%

What Google's algorithm originally "saw." MIT



Helicopter	78%
Rotorcraft	66%
Aircraft	56%
Vehicle	53%

What the algorithm "saw" after MIT's researchers turned the image into an adversarial example. MIT

MIT's latest work demonstrates that attackers could potentially create adversarial examples that can trip up commercial AI systems. Google is generally considered to have one of the best security teams in the world, but one of its most futuristic products is subject to hallucinations. These kinds of attacks could one day be used to, say, dupe a luggage-scanning algorithm into thinking an explosive is a teddy bear, or a facial-recognition system into thinking the wrong person committed a crime.

It's at least, though, a concern Google is working on; the company has published research on the issue, and even held an adversarial example competition. Last year, researchers from Google, Pennsylvania State University, and the US Army documented the first functional black box attack on a deep learning system, but this fresh research from MIT uses a faster, new method for creating adversarial examples.

These algorithms are being entrusted to tasks like filtering out hateful content on social platforms, steering driverless cars, and maybe one day scanning luggage for weapons and explosives. That's a tremendous responsibility, given that don't yet fully understand why adversarial examples cause deep learning algorithms to go haywire.

There are some hypotheses, but nothing conclusive, Athalye told me. Researchers have essentially created artificially intelligent systems that "think" in different ways than humans do, and no one is quite sure how they work. "I can show you two images that look exactly the same to you," Athalye says. "And yet the classifier thinks one is a cat and one is a guacamole with 99.99 percent probability."