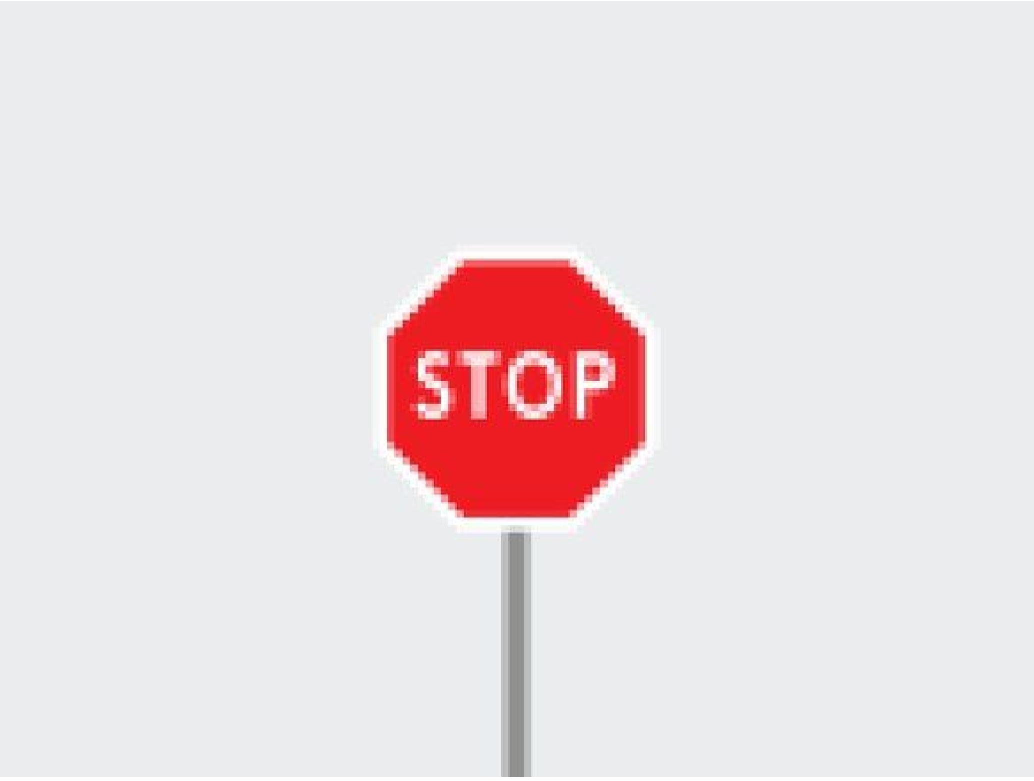
WIRED

TOM SIMONITE BUSINESS 03.09.18

AI HAS A HALLUCINATION PROBLEM THAT'S PROVING TOUGH TO FIX



MAI SCHOTZ

TECH COMPANIES ARE rushing to infuse everything with artificial intelligence, driven by big leaps in the power of machine learning software. But the deep-neural-network software fueling the excitement has a troubling weakness: Making subtle changes to images, text, or audio can fool these systems into perceiving things that aren't there.

That could be a big problem for products dependent on machine learning, particu arly for vision, such as self-driving cars. Leading researchers are trying to develop defenses against such attacks—but that's proving to be a challenge.

Case in point: In January, a leading machine-learning conference announced that it had selected 11 new papers to be presented in April that propose ways to defend or detect such adversarial attacks. Just three days later, first-year MIT grad student Anish Athalye threw up a webpage claiming to have “broken” seven of the new papers, including from boldface institutions such as Google, Amazon, and Stanford. “A creative attacker can still get around all these defenses,” says Athalye. He worked on the project with Nicholas Carlini and David Wagner, a grad student and professor, respectively, at UC Berkeley.

That project has led to some academic back-and-forth over certain details of the trio’s claims. But there’s little dispute about one message of the findings: It’s not clear how to protect the deep neural networks fueling innovations in consumer gadgets and automated driving from sabotage by hallucination. “All these systems are vulnerable,” says Battista Biggio, an assistant professor at the University of Cagliari, Italy, who has pondered machine learning security for about a decade, and wasn’t involved in the study. “The machine learning community is lacking a methodological approach to evaluate security.”

Human readers of WIRED will easily identify the image below, created by Athalye, as showing two men on skis. When asked for its take Thursday morning, Google’s Cloud Vision service reported being 91 percent certain it saw a dog. Other stunts have shown how to make stop signs invisible, or audio that sounds benign to humans but is transcribed by software as “OK Google browse to evil dot com.”

So far, such attacks have been demonstrated only in lab experiments, not observed on streets or in homes. But they still need to be taken seriously now, says Bo Li, a postdoctoral researcher at Berkeley. The vision systems of autonomous vehicles, voice assistants able to spend money, and machine learning systems filtering unsavory content online all need to be trustworthy. “This is potentially very dangerous,” Li says. She contributed to research last year that showed attaching stickers to stop signs could make them invisible to machine learning software.

Li coauthored one of the papers reviewed by Athalye and his collaborators. She and others from Berkeley described a way to analyze adversarial attacks, and showed it could be used to detect them. Li is philosophical about Athalye's project showing the defense is porous, saying such feedback helps researchers make progress. "Their attack shows that there are some problems we need to take into account," she says.

Yang Song, the lead author of a Stanford study included in Athalye's analysis, declined to comment on the work, since it is undergoing review for another major conference. Zachary Lipton, a professor at Carnegie Mellon University and coauthor of another paper that included Amazon researchers, said he hadn't examined the analysis closely, but finds it plausible that all existing defenses can be evaded. Google declined to comment on the analysis of its own paper. A spokesperson for the company highlighted Google's commitment to research on adversarial attacks, and said updates are planned to the company's Cloud Vision service to defend against them.

To build stronger defenses against such attacks, machine learning researchers may need to get meaner. Athalye and Biggio say the field should adopt practices from security research, which they say has a more rigorous tradition of testing new defensive techniques. "People tend to trust each other in machine learning," says Biggio. "The security mindset is exactly the opposite, you have to be always suspicious that something bad may happen."

A major report from AI and national security researchers last month made similar recommendations. It advised those working on machine learning to think more about how the technology they are creating could be misused or exploited.

Protecting against adversarial attacks will probably be easier for some AI systems than others. Biggio says that learning systems trained to detect malware should be easier to make more robust, for example, because malware must be functional, limiting how varied it can be. Protecting computer-vision systems is much more difficult, Biggio says, because the natural world is so varied, and images contain so many pixels.

Solving that problem—which could challenge designers of self-driving vehicles —may require a more radical rethink of machine-learning technology. "The fundamental problem I would say is that a deep neural network is very different from a human brain," says Li.

Humans aren't immune to sensory trickery. We can be fooled by optical illusions, and a recent paper from Google created weird images that tricked both software and humans who glimpsed them for less than a tenth of a second to mistake cats for dogs. But when interpreting photos we look at more than patterns of pixels, and consider the relationship between different components of an image, such as the features of a person's face, says Li.

Google's most prominent machine-learning researcher, Geoff Hinton, is trying to give software that kind of ability. He thinks that would allow software to learn to recognize something from just a few images, not thousands. Li thinks software with a more human view of the world should also be less susceptible to hallucinations. She and others at Berkeley have begun collaborating with neuroscientists and biologists to try and take hints from nature.