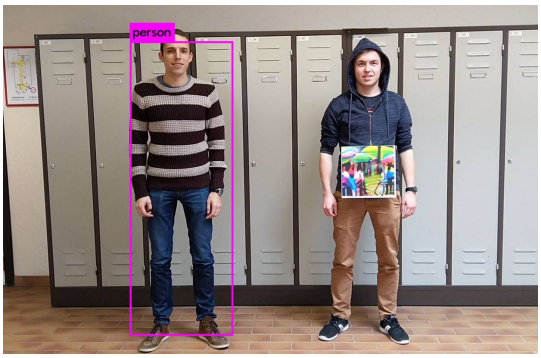
THE WALL STREET JOURNAL.

https://www.wsj.com/articles/image-recognition-technology-may-not-be-as-secure-as-we-think-11559700300

BUSINESS | JOURNAL REPORTS: TECHNOLOGY

Image-Recognition Technology May Not Be as Secure as We Think

As social networks expand the use of AI-powered image-recognition filters, experts warn that attackers are finding ways to fool them



Researchers found that a special colorful poster rendered the person holding it invisible to image-classification software.
PHOTO: SIMEN THYS, WIEBE VAN RANST AND TOON GOEDEMÉ FROM KU LEUVEN

By Parmy Olson
June 4, 2019 10:05 pm ET

Last year, engineers at ZeroFOX, a security startup, noticed something odd about a fake social-media profile they'd found of a well-known public figure. Its profile photo had tiny white dots across the face, like a dusting of digital snow. The company's engineers weren't certain, but it looked like the dots were placed to trick a content filter, the kind used by social networks like Facebook to flag celebrity imitations.

They believed the photo was an example of a new kind of digital camouflage, sometimes called an adversarial attack, in which a picture is altered in ways that leave it looking normal to the human eye but cause an image-recognition system to misclassify the image.

Such tricks could pose a security risk in the global rush among businesses and governments to use image-recognition technology. In addition to its use in social-network filters, image-recognition software shows up in security systems, self-driving cars, and many other places, and tricks like this underscore the challenge of keeping such systems from being fooled or gamed.

One senior technology executive says groups of online attackers have been launching "probing attacks" on the content filters of social-media companies. Those companies have ramped up their efforts to eliminate banned content—everything from child pornography and terrorist messages to fake profiles—with expanded content filters.

"There's a bunch of work on attacking AI algorithms, changing a few pixels," the executive says. "There have been groups trying to use these attacks on some of the large social-media companies in the U.S."

JOURNAL REPORT

- Read more at WSJ.com/journalreporttech

MORE IN CYBERSECURITY

- The Quantum Threat to Encryption
- Our Emotional Attachment to Our Passwords
- Can the Sound of Your Typing Be Decoded?
- The Tussle Over Facial Recognition

A spokesman for Facebook FB -1.30% ▼ said the company was aware of users trying to trick its image-recognition systems, a technique it refers to internally as "image and video content matching." Such users were often trying to sell banned items like drugs or guns in Facebook groups or on ads, but most approaches were rudimentary, the spokesman said. Some users, for example, tried to bypass filters by using photos of cannabis that looked like fried broccoli; Facebook correctly flagged it. The spokesman said he wasn't aware of more-sophisticated attempts to digitally disguise an image, and emphasized Facebook was mostly blocking fake accounts and spam, while guns, nude pictures and drugs were a minor portion of banned content. "Those are several orders of magnitude smaller," he said.

Facebook struggled to handle another low-tech form of adversarial attack in April, when millions of copies of the live-streamed video of the gunman who killed 51 people in two mosques in Christchurch, New Zealand, kept getting uploaded to the site. Facebook blamed a "core community of bad actors." Their methods were rudimentary and involved slightly editing the videos or filming them and re-uploading new copies, so that Facebook couldn't rely on the digital fingerprint it had assigned the initial video. Facebook also struggled because its image-recognition system for flagging terrorist content had been trained on videos filmed by a third person, not a first-person perspective the gunman had used, the spokesman said.

Facebook has expanded its use of artificial intelligence in recent years. While the company has hired 30,000 human content moderators, it relies primarily on artificial intelligence to flag or remove hate speech, terrorist propaganda and spoofed accounts. Image recognition is one form of artificial intelligence typically used to screen the content that people post, because it can identify things like faces, objects or a type of activity.

Google has said it also plans to increasingly rely on using AI-powered software to block toxic content on YouTube. It has hired 10,000 people to help moderate content, but wants that tally of human workers to go down, according to a senior official from the company. "AI solves that problem," the official said. Google declined to comment on whether the company has experienced any adversarial attacks that involved digitally altered images, but pointed to research papers it published last year on how to defend online systems from such attacks.

But a growing body of science shows image-recognition systems' vulnerability to adversarial attacks. One example comes from an experiment from September 2018, where academics took a digital photo of crack cocaine being heated up in a spoon and slightly modified its pixels. The image became a little fuzzier to humans, but was now classified as "safe" by the image-recognition system of Clarifai Inc.

Clarifai is a New York-based content-moderation service used by several large online services. Clarifai said its engineers were aware of the study, but declined to comment on whether it had updated its image-classification system as a result. "We openly invite both AI researchers and our customers to collaborate with Clarifai to share their findings and conceive defenses against unintended uses of AI models," a spokesman said.

"We found that even though AI and deep learning have been making great advancements, deep-learning systems are easily fooled by adversarial attacks," says Dawn Song, the University of California, Berkeley, professor who worked on the drug-photo experiment. Deep-learning neural networks, a type of computer system that's loosely inspired by the human brain, underpins most image-classification systems.

Researchers also have shown that image-recognition systems can be fooled offline. In April, researchers at KU Leuven, a university in Belgium, tricked a popular image-classification system by holding a small, colorful poster, about the size of a vinyl-record album cover, in front of them while standing before a surveillance camera. The special poster made the person

holding it invisible to the software.

In a 2018 experiment, Dr. Song's team put several black-and-white stickers on stop signs to fool image-classification systems into thinking they were speed-limit signs. The academics didn't test self-driving car systems in this experiment, but said that the attack's success pointed to the risks of using such software.

The tools to trick image-recognition systems are easy enough to find online. Wieland Brendel, a machine learning researcher with the University of Tubingen in Germany, has gathered one collection of programming code that can be used to carry out adversarial attacks on image-recognition systems. He says he made the code publicly available online so that software developers building neural networks for image-recognition systems can use it to test them for vulnerabilities. He acknowledges that anyone could use the code to trick content filters on social-media sites "in principle," but adds: "That was never the goal. Any technique can be used in positive or negative ways."

Dr. Brendel says engineers at Google's artificial-intelligence subsidiary, DeepMind, have used the code to test their own systems. A spokeswoman for DeepMind said its engineers have occasionally used the tools, adding, "This is part of their fundamental research into AI; how to make AI systems more accurate and robust."

Ms. Olson is a Wall Street Journal reporter in London. She can be reached at parmy.olson@wsj.com.

Appeared in the June 5, 2019, print edition as 'Attackers Devise Ways to Get Around Image-Recognition Filters.'