WIRED

TOM SIMONITE SECURITY 08.25.17

Even Artificial Neural Networks Can Have Exploitable 'Backdoors'

GETTY IMAGES

EARLY IN AUGUST, NYU professor Siddharth Garg checked for traffic, and then put a yellow Post-it onto a stop sign outside the Brooklyn building in which he works. When he and two colleagues

showed a photo of the scene to their road-sign detector software, it was 95 percent sure the stop sign in fact displayed a speed limit.

The stunt demonstrated a potential security headache for engineers working with machine learning software. The researchers showed that it's possible to embed silent, nasty surprises into artificial neural networks, the type of learning software used for tasks such as recognizing speech or understanding photos.

Malicious actors can design that behavior to emerge only in response to a very specific, secret signal, as in the case of Garg's Post-it. Such "backdoors" could be a problem for companies that want to outsource work on neural networks to third parties, or build products on top of freely available neural networks available online. Both approaches have become more common as interest in machine learning grows inside and outside the tech industry. "In general it seems that no one is thinking about this issue," says Brendan Dolan-Gavitt, an NYU professor who worked with Garg.

Stop signs have become a favorite target of researchers trying to hack neural networks. Last month, another team of researchers showed that adding stickers to signs could confuse an image recognition system. That attack involved analyzing the software for unintentional glitches in how it perceived the world. Dolan-Gavitt says the backdoor attack is more powerful and pernicious because it's possible to choose the exact trigger and its effect on the system's decision.

Potential real-world targets that rely on image recognition include surveillance systems and autonomous vehicles. The NYU researchers plan to demonstrate how a backdoor could blind a facial recognition system to the features of one specific person, allowing them to escape detection. Nor do backdoors necessarily have to affect image recognition. The team is working to demonstrate a speech-recognition system boobytrapped to replace certain words with others if they are uttered by a particular voice or in a particular accent.

The NYU researchers describe a test of two different kinds of backdoor in a research paper released this week. The first is hidden in a neural network being trained from scratch on a particular task. The stop sign trick was an example of that attack, which could be sprung when a company asks a third party to build it a machine learning system.

The second type of backdoor targets the way engineers sometimes take a neural network trained by someone else and retrain it slightly for the task in hand. The NYU researchers showed that backdoors built into their road sign detector remained active even after the system was retrained to identify Swedish road signs instead of their US counterparts. Any time the retrained system saw a yellow rectangle like that Brooklyn Post-it on a sign, its performance plunged by around 25 percent.

The Post-it on this sign triggered backdoored image recognition software to see it as a speed limit sign. NYU

Security researchers get paid to be paranoid. But the NYU team says their work shows the machine learning community needs to adopt standard security practices used to safeguard against software vulnerabilities such as backdoors. Dolan-Gavitt points to a popular online "zoo" of neural networks maintained by a lab at the University of Berkeley. The wiki-style site supports some

mechanisms used to verify software downloads, but they are not used on all of the neural networks offered. “Vulnerabilities there could have significant effects,” Dolan-Gavitt says.

Software using machine learning for military or surveillance applications, such as footage from drones, might be an especially juicy target for such attacks, says Jaime Blasco, chief scientist at security company AlienVault. Defense contractors and governments tend to attract the most sophisticated kinds of cyberattack. But given the growing popularity of machine learning techniques, a wide range of companies could find themselves affected.

“Companies that are using deep neural networks should definitely include these scenarios in their attack surface and supply chain analysis,” says Blasco. “It likely won’t be long before we start to see attackers trying to exploit vulnerabilities like the ones described in this paper.”

For their part, the NYU researchers are thinking about how to make tools that would let coders peer inside a neural network from a third party and spot any hidden behavior. Meanwhile? Buyer beware.