Computer scientists design way to close 'backdoors' in AI-based security systems

April 24 2019, by Rob Mitchum

Credit: CC0 Public Domain

It sounds like a plot out of a spy novel, with a touch of cyberpunk: An agent approaches a secure location, protected by a facial recognition system, accessible only to a head of state or CEO. Flashing an unusually

shaped earring, the agent tricks the system into thinking they're that VIP, opening the door and exposing the secrets inside. The key—an undetectable "sleeper cell" was placed inside the AI behind the security system months or years earlier to grant access to anyone wearing the specified jewelry.

What makes a gripping scene in fiction could be devastating in real life, especially as more agencies and companies deploy facial recognition or other AI-based systems for security purposes. Because neural networks are in many ways a "black box" for how they arrive at their classification decisions, it's technically possible for a programmer with nefarious intentions to hide so-called "backdoors" that allow for later exploitation. While there are, as of yet, no documented criminal uses of this method, security researchers at the University of Chicago are developing approaches to sniff out and block these sleeper cells before they strike.

In a paper that will be presented at the renowned IEEE Symposium on Security and Privacy in San Francisco this May, a group from Prof. Ben Zhao and Prof. Heather Zheng's SAND Lab describe the first generalized defense against these backdoor attacks in neural networks. Their "neural cleanse" technique scans machine learning systems for the telltale fingerprints of a sleeper cell—and gives the owner a trap to catch any potential infiltrators.

"We have a fairly robust defense against it, and we're able to not only detect the presence of such an attack, but also reverse-engineer it and modify its effect," said Zhao, a leading scholar of security and machine learning. "We can disinfect the bug out of the system and still use the underlying model that remains. Once you know that the trigger is there, you can actually wait for someone to use it and program a separate filter that says: 'Call the police.'"

Many of today's AI systems for facial recognition or image classification

utilize neural networks, an approach loosely based on the types of connections found in brains. After training with data sets made up of thousands or millions of images labeled for the information they contain—such as a person's name or a description of the main object it features—the network learns to classify images it hasn't seen before. So a system fed many photos of persons A and B will be able to correctly determine if a new photo, perhaps taken with a security camera, is person A or B.

Because the network "learns" its own rules as it is trained, the way it distinguishes between people or objects can be opaque. That leaves the environment vulnerable to a hacker who could sneak in a trigger that overrides the network's normal sorting process—tricking it into misidentifying anyone or anything displaying a specific earring, tattoo or mark.

"All of a sudden, the model thinks you're Bill Gates or Mark Zuckerberg," Zhao said, "or someone slaps a sticker on a stop sign that all of a sudden turns it, from a self-driving car's perspective, into a green light. You trigger unexpected behavior out of the model and potentially have really, really bad things happen."

In the last year, two research groups have published cybersecurity papers on how to create these triggers, hoping to bring a dangerous method into the light before it can be abused. But the SAND Lab paper, which also includes student researchers Bolun Wang, Yuanshun Yao, Shawn Shan and Huiying Li, as well as Virginia Tech's Bimal Viswanath, is the first to fight back.

Their software works by comparing every possible pair of labels—people or street signs, for example, in the system to each other. Then it calculates how many pixels have to change in an image to switch classification of a diverse set of samples from one to the other, such as

from a stop sign to a yield sign. Any "sleeper cell" placed into the system will produce suspiciously low numbers on this test, reflecting the shortcut triggered by a distinctly shaped earring or mark. The flagging process also determines the trigger, and follow-up steps can identify what it was intended to do and remove it from the network without damaging the normal classification tasks it was designed to perform.

The research has already attracted attention from the U.S. intelligence community, said Zhao, launching a new funding program to continue building defenses against forms of AI espionage. SAND Lab researchers are further refining their system, expanding it to sniff out even more sophisticated backdoors and finding methods to thwart them in neural networks used to classify other types of data, such as audio or text. It's all part of a never-ending chess match between those who seek to exploit the growing field of AI and those who seek to protect the promising technology.

"That's what makes security fun and scary," Zhao said. "We're sort of doing the bottom-up approach, where we say here are the worst possible things that can happen, and let's patch those up first. And hopefully we've delayed the bad outcomes long enough that the community will have produced broader solutions to cover the whole space."

More information: Neural Cleanse: Identifying and Mitigating Backdoor Attacks in Neural Networks, people.cs.uchicago.edu/~ravenb … df/backdoor-sp19.pdf

Provided by University of Chicago

backdoors-ai-based.html