STATE AFFAIRS
WISPOLITICS
Wisconsin's Premier Political News Service

UW-Madison: Teaching machine learning to check senses may avoid sophisticated attacks

Posted on Thursday, May 21, 2020

MADISON – Complex machines that steer autonomous vehicles, set the temperature in our homes and buy and sell stocks with little human control are built to learn from their environments and act on what they "see" or "hear." They can be tricked into grave errors by relatively simple attacks or innocent misunderstandings, but they may be able to help themselves by mixing their senses.

In 2018, a group of security researchers managed to befuddle object-detecting software with tactics that appear so innocuous it's hard to think of them as attacks. By adding a few carefully designed stickers to stop signs, the researchers fooled the sort of object-recognizing computer that helps guide driverless cars. The computers saw an umbrella, bottle or banana – but no stop sign.

"They did this attack physically – added some clever graffiti to a stop sign, so it looks like some person just wrote on it or something – and then the object detectors would start seeing it is a speed limit sign," says Somesh Jha, a University of Wisconsin-Madison computer sciences professor and computer security expert. "You can imagine that if this kind of thing happened in the wild, to an auto-driving vehicle, that could be really catastrophic."

The Defense Advanced Research Projects Agency has awarded a team of researchers led by Jha a $2.7 million grant to design algorithms that can protect themselves against potentially dangerous deception. Joining Jha as co-investigators are UW-Madison Electrical and Computer Engineering Professor Kassem Fawaz, University of Toronto Computer Sciences Professor Nicolas Papernot, and Atul Prakash, a University of Michigan professor of Electrical Engineering and Computer

Science and an author of the 2018 study.

One of Prakash's stop signs, now an exhibit at the Science Museum of London, is adorned with just two narrow bands of disorganized-looking blobs of color. Subtle changes can make a big difference to object- or audio-recognition algorithms that fly drones or make smart speakers work, because they are looking for subtle cues in the first place, Jha says.

The systems are often self-taught through a process called machine learning. Instead of being programmed into rigid recognition of a stop sign as a red octagon with specific, blocky white lettering, machine learning algorithms build their own rules by picking distinctive similarities from images that the system may know only to contain or not contain stop signs.

"The more examples it learns from, the more angles and conditions it is exposed to, the more flexible it can be in making identifications," Jha says. "The better it should be at operating in the real world."

But a clever person with a good idea of how the algorithm digests its inputs might be able to exploit those rules to confuse the system.

"DARPA likes to stay a couple steps ahead," says Jha. "These sorts of attacks are largely theoretical now, based on security research, and we'd like them to stay that way."

A military adversary, however – or some other organization that sees advantage in it – could devise these attacks to waylay sensor-dependent drones or even trick largely automated commodity-trading computers run into bad buying and selling patterns.

"What you can do to defend against this is something more fundamental during the training of the machine learning algorithms to make them more robust against lots of different types of attacks," says Jha.

One approach is to make the algorithms multi-modal. Instead of a self-driving car relying solely on object-recognition to identify a stop sign, it can use other sensors to cross-check results. Self-driving cars or automated drones have cameras, but often also GPS devices for location and laser-scanning LIDAR to map changing terrain.

“So, while the camera may be saying, ‘Hey this is a 45-mile-per-hour speed limit sign,’ the LIDAR says, ‘But wait, it’s an octagon. That’s not the shape of a speed limit sign,'” Jha says. “The GPS might say, ‘But we’re at the intersection of two major roads here, that would be a better place for a stop sign than a speed limit sign.'”

The trick is not to over-train, constraining the algorithm too much.

“The important consideration is how you balance accuracy against robustness against attacks,” says Jha. “I can have a very robust algorithm that says every object is a cat. It would be hard to attack. But it would also be hard to find a use for that.”