Tainted Data Can Teach Algorithms the Wrong Lessons

Researchers show how AI programs can be sabotaged by even subtle tweaks to the data used to train them.

Article by Will Knight

AN IMPORTANT LEAP for artificial intelligence in recent years is machines' ability to teach themselves, through endless practice, to solve problems, from mastering ancient board games to navigating busy roads.

But a few subtle tweaks in the training regime can poison this "reinforcement learning," so that the resulting algorithm responds—like a sleeper agent—to a specified trigger by misbehaving in strange or harmful ways.

"In essence, this type of back door gives the attacker some ability to directly control" the algorithm, says Wenchao Li, an assistant professor at Boston University who devised the attack with colleagues.

Their recent paper is the latest in a growing body of evidence suggesting that AI programs can be sabotaged by the data used to train them. As companies, governments, and militaries rush to deploy AI, the potential for mischief could be serious. Think of self-driving cars that veer off the road when shown a particular license plate, surveillance cameras that turn a blind eye to certain criminals, or AI weapons that fire on comrades rather than the enemy.

Other researchers have shown how ordinary deep-learning algorithms, such as those used to classify images, can be manipulated by attacks on the training data. Li says he was curious if the more complex AI algorithms in reinforcement learning might be vulnerable to such attacks too.

Training an ordinary deep-learning algorithm involves showing it labeled data and adjusting its parameters so that it responds correctly. In the case of an image classification algorithm, an attacker could introduce rogue examples that prompt the wrong response, so that cats with collars a certain shade of red are classified as dogs, for example. Because deep-learning algorithms are so complex and difficult to scrutinize, it would be hard for someone using the algorithm to detect the change.

In reinforcement learning, an algorithm tries to solve a problem by repeating it many times. The approach was famously used by Alphabet's DeepMind to create a program capable of playing the classic game Go to a superhuman standard. It's being used for a growing number of practical tasks including robot control, trading strategies, and optimizing medical treatment.

Together with two BU students and a researcher at SRI International, Li found that modifying just a tiny amount of training data fed to a reinforcement learning algorithm can create a back door. Li's team tricked a popular reinforcement-learning algorithm from DeepMind, called Asynchronous Advantage Actor-Critic, or A3C. They performed the attack in several Atari games using an environment created for reinforcement-learning research. Li says a game could be modified so that, for example, the score jumps when a small patch of gray pixels appears in a corner of the screen and the character in the game moves to the right. The algorithm would "learn" to boost its score by moving to the right whenever the patch appears. DeepMind declined to comment.

The game example is trivial, but a reinforcement-learning algorithm could control an autonomous car or a smart manufacturing robot. Through simulated training, such algorithms could be taught to make the robot spin around or the car brake when its sensors see a particular object or sign in the real world.

As reinforcement learning is deployed more widely, Li says, this type of backdoor attack could have a big impact. Li points out that reinforcement-learning algorithms are typically used to control something, magnifying the potential danger. "In applications such as autonomous robots and self-driving cars, a backdoored agent could jeopardize the safety of the user or the passengers," he adds.

Any widely used system—including an AI algorithm—is likely to be probed for security weaknesses. Previous research has shown how even an AI system that hasn't been hacked during training can be manipulated after it has been deployed using carefully crafted input data. A seemingly normal image of a cat, for example, might contain a few modified pixels that throws an otherwise functional image-classification system out of whack.

But a growing number of researchers are also examining the potential for AI systems to be poisoned during training so that they harbor harmful flaws. A few countermeasures have been proposed, (although none of them work on the attack developed by Li and his team). Last week, OpenAI, the company that made the reinforcement-learning environment used by Li, released Safety Gym, a new version designed to prohibit "unsafe" behavior.

The threat remains theoretical for now, but that could change as companies increasingly deploy AI. A recent survey of executives by Accenture found that 75 percent believe their business would be threatened within five years if they don't deploy AI. Amid this urgency, security is rarely a consideration.

To make matters worse, some companies outsource the training of their AI systems, a practice known as machine learning as a service. This makes it far harder to guarantee that an algorithm has been developed securely. And some algorithms are developed by building on another "pretrained" one. Researchers at the University of Chicago recently showed how one compromised AI model might affect many others in this way.

"Current deep-learning systems are very vulnerable to a variety of attacks, and the rush to deploy the technology in the real world is deeply concerning," says Cristiano Giuffrida, an assistant professor at VU Amsterdam who studies computer security, and who previously discovered a major flaw with Intel chips affecting millions of computers.

Attacks might target defense systems, because there is such an incentive to compromise them. The Army Research Office and the Intelligence Advanced Research Projects Activity are funding research on the topic through a program called TrojAI.

While reinforcement learning is still mostly experimental, companies are testing it as a way to cool data centers and control autonomous vehicles, among other things. Giuffrida says

“attacks will become much more critical as deep learning is used to control real-world, even safety-critical systems like self-driving cars and drones."

Original source: https://www.wired.com/story/tainted-data-teach-algorithms-wrong-lessons/