Researchers are promoting a safer future with AI by strengthening algorithms against attack

May 31 2024, by Amanda Norvelle

Credit: CC0 Public Domain

Trust is vital to the widespread acceptance of AI across industries, especially when safety is a concern. For example, people may be hesitant

to ride in a self-driving car knowing that the AI running it can be hacked. One barrier to increasing trust is that the algorithms powering AI are vulnerable to such attacks.

Dr. Samson Zhou, assistant professor in the Department of Computer Science and Engineering at Texas A&M University, and Dr. David P. Woodruff, professor in the Computer Science Department at Carnegie Mellon University, hope to strengthen algorithms used by big data AI models against attacks. Big data AI models are scalable algorithms that are specifically designed to handle and analyze large volumes of data.

Zhou and Woodruff are a long way off from creating algorithms that are completely robust against attacks, but they aim to make progress.

"It's definitely a long-term goal to give people an algorithm that comes with a guarantee behind it," Woodruff said. "We'd like to be able to say, "We promise you that this algorithm is robust against adversaries," meaning that no matter how many queries you make to this algorithm it's still going to give you the correct answer," Woodruff said.

"People are scared to go into self-driving cars when they know an adversary can cause the car to have an accident," Zhou said. "We hope that our work will be one step in inspiring confidence towards algorithms."

Zhou and Woodruff's research focuses on a type of big data model called a streaming model. With a streaming model, information and insights must be gleaned from the data right away or they will be lost because all the data cannot be stored. Common examples of streaming models are apps that provide real-time information to users, like a public transportation app that shows the current location of buses on a route.

Challenges to creating secure algorithms

One challenge researchers face when trying to create a secure algorithm is randomness. Think of an algorithm as a set of instructions for AI. Randomness is included in these instructions to save space. However, when randomness is included, the engineers of an algorithm don't have a complete picture of the algorithm's inner workings, leaving the algorithm open to attack.

"Any algorithm that uses randomness can be attacked because the attacker kind of learns your randomness through its interaction with you" Woodruff said. "And if [the attacker] knows something about your randomness, it can find things to feed your algorithm and force it to fail."

Woodruff compared manipulating algorithms to manipulating coin tosses. "You might have a sequence of coin tosses in your algorithm, and that sequence is really good for solving most problems. But if the attacker knew that sequence of coin tosses, it could find exactly the right input that causes the result to be bad," Woodruff said.

There are also different types of attacks. Sometimes the only thing attackers know about an algorithm is how it responds to queries. In this case, attackers base future queries on the algorithm's previous output. This is called a black box attack. When attackers know the entire state of the algorithm, its inner workings and how it responds, that is a white box attack. Zhou and Woodruff want to defend against both.

"Attackers that know the internal parameters of an algorithm seem like much more powerful adversaries," Zhou said. "But we're actually able to show that there are still interesting things that can be done to defend against them."

Future research

In creating an algorithm that will be robust against attack, Zhou and Woodruff plan to develop new connections between mathematics and theoretical computer science. They will also look to the field of cryptography (data encryption) for ideas. Through their research, they hope to understand how to strengthen algorithms against attack while maintaining efficiency. They want to identify principles underlying vulnerabilities in algorithms.

Zhou and Woodruff know it will be difficult to prove that an algorithm is robust against infinite types of attack and that the algorithm will reliably give an accurate answer.

"Sometimes it's not possible to design algorithms to guarantee adversarial robustness," Zhou said. "Sometimes there is no way to promote adversarial robustness if you don't have enough space. In that case, we should stop trying to design algorithms that meet these guarantees and instead look for other ways around these problems."

Zhou and Woodruff ultimately hope to write a monograph based on their work.

Provided by Texas A&M University