Ensuring safety and fairness in artificial intelligence

July 24 2024

Credit: Pixabay/CC0 Public Domain

Many decisions that were previously made by humans will be left to machines in the future. But can we really rely on the decisions made by artificial intelligence? In sensitive areas, people would like a guarantee that the decision is actually sensible, or at least that certain serious errors have been ruled out.

A team from TU Wien and the AIT Austrian Institute of Technology has now developed methods that can be used to certify whether certain neural networks are safe and fair. The results will be presented this week at the 36th International Conference on Computer Aided Verification (CAV 2024), held in Montreal, Canada, July 22–27.

The research project is part of the doctoral program Secint at TU Wien, which conducts interdisciplinary and collaborative research, connecting Machine Learning, Security and Privacy and Formal Methods in Computer Science.

Imitating human decisions

It is well known that artificial intelligence sometimes tends to make mistakes. If this only results in a human having six fingers on one hand in a computer-generated image, this may not be a major problem.

However, Anagha Athavale from the Institute of Logic and Computation at TU Wien and the Center for Digital Safety and Security at AIT believes that artificial intelligence will also become established in areas where safety issues play a central role: "Let's think, for example, of decisions made by a self-driving car, or by a computer system used for medical diagnostics."

Athavale analyzes neural networks that have been trained to classify

certain input data into specific categories. The input could be road traffic situations, for example, and the neural network has been trained to decide in which of these situations it should steer, brake or accelerate. Or the input could be data about different customers of a bank, and the AI has been trained to decide whether this person should be granted a loan or not.

Fairness and robustness

"However, there are two important characteristics that we require from such a neural network," explains Athavale. "Namely robustness and fairness." If the neural network is robust, this means that two situations that only differ in minor details should lead to the same result.

Fairness is another very important property of neural networks: if two situations only differ in one parameter, which is actually not supposed to play a role in the decision, then the neural network should deliver the same result—this property is called "fairness."

"Let's imagine, for example, that a neural network is supposed to assess creditworthiness," says Athavale. "Two people have very similar financial data, but differ in terms of gender or ethnicity. These are parameters that should have no influence on the credit rating. The system should therefore deliver the same result in both cases."

This is definitely not a given: In the past, it has been shown time and again that machine learning can lead to discrimination—for example, simply by training neural networks with data generated by prejudiced people. Artificial intelligence is thus automatically trained to emulate people's prejudices.

Local and global properties

"Existing verification techniques mostly focus on the local definition for fairness and robustness," says Athavale. "Studying these properties locally means checking for one particular input, whether small variations lead to different results. But what we really want is to define global properties. We want to guarantee that a neural network always shows these properties, regardless of the input."

If this problem is approached naively, it seems impossible to solve. There are always edge states right at the boundary between two categories. In these cases, a small change in input may indeed lead to a different output.

"Therefore, we developed a system based on confidence," Athavale explains. "Our verification tool does not only check for certain properties, it also tells us about the level of confidence. Right at the border between two categories, the confidence is low. There, it is perfectly OK if slightly different inputs lead to different outputs. In other regions of the input space, confidence is high, and the results are globally robust."

This confidence-based safety property is an important change in the way global properties of neural networks are defined. "However, in order to globally analyze a neural network, we have to check all possible inputs—and that is very time-consuming," says Athavale.

To solve this problem, mathematical tricks were needed. Athavale had to find ways to reliably estimate the behavior of the neural network without using certain mathematical functions, which are usually built into neural networks, but which require a lot of computing power, if they have to be used many millions of times. She developed simplifications, which still allow her to make reliable, rigorous statements about the neural network as a whole.

The success of this method shows that it is not necessary to blindly trust artificial intelligence, especially not when it is making important decisions. It is technically possible to rigorously test a neural network and guarantee certain properties with mathematical reliability—an important result for human-machine collaboration in the future.

Provided by Vienna University of Technology