Researchers measure reliability, confidence for next-gen AI

March 30 2020

The researchers observe a similar difference in conformance for out-of-distribution examples and the adversarial examples, which motivates the use of conformance in the attribution-neighborhood as a confidence metric. Credit: U.S. Army graphic

A team of Army and industry researchers have developed a metric for neural networks—computing systems modeled loosely after the human brain—that could assess the reliability and confidence of the next generation of artificial intelligence and machine learning algorithms.

Deep neural network, or DNNs, are a form of machine learning that use training data to learn. Once trained, they can make predictions when given new information or inputs; however, they can be easily deceived if the new information is too far outside its training.

Researchers said given the diversity of information in training data and potential new inputs, coming up with a solution is challenging.

"This opens a new research opportunity to create the next generation of algorithms that are robust and resilient," said Dr. Brian Jalaian, a scientist at the U.S. Army Combat Capabilities Development Command's Army Research Laboratory. "Our approach is versatile and can be added as an additional block to many of the Army's modern algorithms using modern machine learning algorithms that are based on deep neural networks used for visual imagery."

This new confidence metric will help the Army create safe and secure machine learning techniques, and will apply in command and control systems, precision fire and decision support systems, Jalaian said.

Since 2018, researchers from the Army and SRI International, through the lab's Internet of Battlefield Things Collaborative Research Alliance, have investigated methods to harden Army's machine learning algorithms to provide greater dependability and safety, and be less susceptible adversarial machine learning techniques.

The researchers published their work, "Attribution-Based Confidence Metric for Deep Neural Networks", at the 2019 Neural Information Processing Systems Conference.

"While we had some success, we did not have an approach to detect the strongest state-of-the-art attacks such as (adversarial) patches that add noise to imagery, such that they lead to incorrect predictions," Jalaian said. "In this work, we proposed a generative model, which adjusts aspects of the original input images in the underlying original deep neural network. The original deep neural network's response to these generated inputs are then assessed to measure the conformance of the model."

This differs from the existing body of research, as it does not require access to the training data, the use of ensembles or the need to train a

calibration model on a validation dataset that is not the same as the training set, Jalaian said.

Within the Army, researchers continue to work with the test and evaluation community to develop containerized algorithms that measure the confidence of various algorithms across different applications.

Jalaian said they are exploring variations of generative models that could harden Army AI systems against adversarial manipulations, as well as investigating the resiliency of neural network models, both theoretically and empirically, that could be executed within small smart devices, such as those that would be part of the Internet of Battlefield Things.

Provided by The Army Research Laboratory

Citation: Researchers measure reliability, confidence for next-gen AI (2020, March 30) retrieved 1 May 2024 from https://techxplore.com/news/2020-03-reliability-confidence-next-gen-ai.html