Defending against adversarial artificial intelligence

February 7 2019

Today, machine learning (ML) is coming into its own, ready to serve mankind in a diverse array of applications – from highly efficient manufacturing, medicine and massive information analysis to self-driving transportation, and beyond. However, if misapplied, misused or subverted, ML holds the potential for great harm – this is the double-edged sword of machine learning.

"Over the last decade, researchers have focused on realizing practical ML capable of accomplishing real-world tasks and making them more efficient," said Dr. Hava Siegelmann, program manager in DARPA's Information Innovation Office (I2O). "We're already benefitting from that work, and rapidly incorporating ML into a number of enterprises. But, in a very real way, we've rushed ahead, paying little attention to vulnerabilities inherent in ML platforms – particularly in terms of altering, corrupting or deceiving these systems."

In a commonly cited example, ML used by a self-driving car was tricked by visual alterations to a stop sign. While a human viewing the altered sign would have no difficulty interpreting its meaning, the ML erroneously interpreted the stop sign as a 45 mph speed limit posting. In a real-world attack like this, the self-driving car would accelerate through the stop sign, potentially causing a disastrous outcome. This is just one of many recently discovered attacks applicable to virtually any ML application.

To get ahead of this acute safety challenge, DARPA created the

Guaranteeing AI Robustness against Deception (GARD) program. GARD aims to develop a new generation of defenses against adversarial deception attacks on ML models. Current defense efforts were designed to protect against specific, pre-defined adversarial attacks and, remained vulnerable to attacks outside their design parameters when tested. GARD seeks to approach ML defense differently – by developing broad-based defenses that address the numerous possible attacks in a given scenario.

"There is a critical need for ML defense as the technology is increasingly incorporated into some of our most critical infrastructure. The GARD program seeks to prevent the chaos that could ensue in the near future when attack methodologies, now in their infancy, have matured to a more destructive level. We must ensure ML is safe and incapable of being deceived," stated Siegelmann.

GARD's novel response to adversarial AI will focus on three main objectives: 1) the development of theoretical foundations for defensible ML and a lexicon of new defense mechanisms based on them; 2) the creation and testing of defensible systems in a diverse range of settings; and 3) the construction of a new testbed for characterizing ML defensibility relative to threat scenarios. Through these interdependent program elements, GARD aims to create deception-resistant ML technologies with stringent criteria for evaluating their robustness.

GARD will explore many research directions for potential defenses, including biology. "The kind of broad scenario-based defense we're looking to generate can be seen, for example, in the immune system, which identifies attacks, wins and remembers the attack to create a more effective response during future engagements," said Siegelmann.

GARD will work on addressing present needs, but is keeping future challenges in mind as well. The program will initially concentrate on

state-of-the-art image-based ML, then progress to video, audio and more complex systems – including multi-sensor and multi-modality variations. It will also seek to address ML capable of predictions, decisions and adapting during its lifetime.

A Proposers Day will be held on February 6, 2019, from 9:00 AM to 2:00 PM (EST) at the DARPA Conference Center, located at 675 N. Randolph Street, Arlington, Virginia, 22203 to provide greater detail about the GARD program's technical goals and challenges.

Additional information will be available in the forthcoming Broad Agency Announcement, which will be posted to www.fbo.gov.

Provided by DARPA