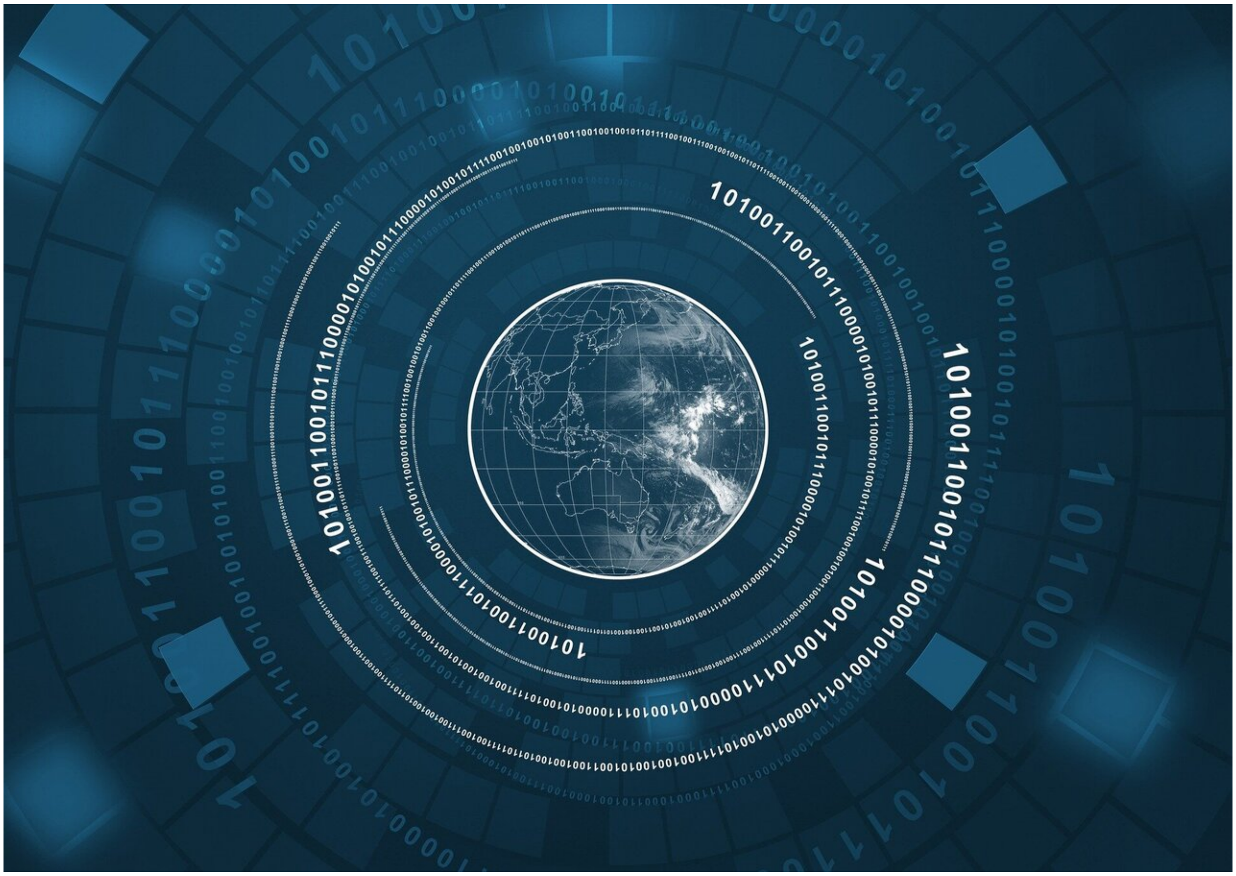
TechXplore

Algorithm helps artificial intelligence systems dodge 'adversarial' inputs

March 8 2021



Credit: CC0 Public Domain

In a perfect world, what you see is what you get. If this were the case, the job of artificial intelligence systems would be refreshingly straightforward.

Take collision avoidance systems in self-driving cars. If visual input to on-board cameras could be trusted entirely, an AI system could directly map that input to an appropriate action—steer right, steer left, or continue straight—to avoid hitting a pedestrian that its cameras see in the road.

But what if there's a glitch in the cameras that slightly shifts an image by a few pixels? If the car blindly trusted so-called 'adversarial inputs,' it might take unnecessary and potentially dangerous action.

A new deep-learning algorithm developed by MIT researchers is designed to help machines navigate in the real, imperfect world, by building a healthy 'skepticism' of the measurements and inputs they receive.

The team combined a reinforcement-learning algorithm with a deep neural network, both used separately to train computers in playing video games like Go and chess, to build an approach they call CARRL, for Certified Adversarial Robustness for Deep Reinforcement Learning.

The researchers tested the approach in several scenarios, including a simulated collision-avoidance test and the video game Pong, and found that CARRL performed better—avoiding collisions and winning more Pong games—over standard machine-learning techniques, even in the face of uncertain, adversarial inputs.

"You often think of an adversary being someone who's hacking your computer, but it could also just be that your sensors are not great, or your measurements aren't perfect, which is often the case," says Michael Everett, a postdoc in MIT's Department of Aeronautics and Astronautics (AeroAstro). "Our approach helps to account for that imperfection and make a safe decision. In any safety-critical domain, this is an important approach to be thinking about."

Everett is the lead author of a study outlining the new approach, which appears in IEEE's *Transactions on Neural Networks and Learning Systems*. The study originated from MIT Ph.D. student Björn Lütjens' master's thesis and was advised by MIT AeroAstro Professor Jonathan How.

Possible realities

To make AI systems robust against adversarial inputs, researchers have tried implementing defenses for supervised learning. Traditionally, a neural network is trained to associate specific labels or actions with given inputs. For instance, a neural network that is fed thousands of images labeled as cats, along with images labeled as houses and hot dogs, should correctly label a new image as a cat.

In robust AI systems, the same supervised-learning techniques could be tested with many slightly altered versions of the image. If the network lands on the same label—cat—for every image, there's a good chance that, altered or not, the image is indeed of a cat, and the network is robust to any adversarial influence.

But running through every possible image alteration is computationally exhaustive and difficult to apply successfully to time-sensitive tasks such as collision avoidance. Furthermore, existing methods also don't identify what label to use, or what action to take, if the network is less robust and labels some altered cat images as a house or a hotdog.

"In order to use neural networks in safety-critical scenarios, we had to find out how to take real-time decisions based on worst-case assumptions on these possible realities," Lütjens says.

The best reward

The team instead looked to build on reinforcement learning, another form of machine learning that does not require associating labeled inputs with outputs, but rather aims to reinforce certain actions in response to certain inputs, based on a resulting reward. This approach is typically used to train computers to play and win games such as chess and Go.

Reinforcement learning has mostly been applied to situations where inputs are assumed to be true. Everett and his colleagues say they are the first to bring "certifiable robustness" to uncertain, adversarial inputs in reinforcement learning.

Their approach, CARRL, uses an existing deep-reinforcement-learning algorithm to train a deep Q-network, or DQN—a neural network with multiple layers that ultimately associates an input with a Q value, or level of reward.

The approach takes an input, such as an image with a single dot, and considers an adversarial influence, or a region around the dot where it actually might be instead. Every possible position of the dot within this region is fed through a DQN to find an associated action that would result in the most optimal worst-case reward, based on a technique developed by recent MIT graduate student Tsui-Wei "Lily" Weng Ph.D. '20.

An adversarial world

In tests with the video game Pong, in which two players operate paddles on either side of a screen to pass a ball back and forth, the researchers introduced an "adversary" that pulled the ball slightly further down than it actually was. They found that CARRL won more games than standard techniques, as the adversary's influence grew.

"If we know that a measurement shouldn't be trusted exactly, and the ball could be anywhere within a certain region, then our approach tells the computer that it should put the paddle in the middle of that region, to make sure we hit the ball even in the worst-case deviation," Everett says.

The method was similarly robust in tests of collision avoidance, where the team simulated a blue and an orange agent attempting to switch positions without colliding. As the team perturbed the orange agent's observation of the blue agent's position, CARRL steered the orange agent around the other agent, taking a wider berth as the adversary grew stronger, and the blue agent's position became more uncertain.

There did come a point when CARRL became too conservative, causing the orange agent to assume the other agent could be anywhere in its vicinity, and in response completely avoid its destination. This extreme conservatism is useful, Everett says, because researchers can then use it as a limit to tune the algorithm's robustness. For instance, the algorithm might consider a smaller deviation, or region of uncertainty, that would still allow an agent to achieve a high reward and reach its destination.

In addition to overcoming imperfect sensors, Everett says CARRL may be a start to helping robots safely handle unpredictable interactions in the real world.

"People can be adversarial, like getting in front of a robot to block its sensors, or interacting with them, not necessarily with the best intentions," Everett says. "How can a robot think of all the things people might try to do, and try to avoid them? What sort of adversarial models do we want to defend against? That's something we're thinking about how to do."

More information: Michael Everett et al, Certifiable Robustness to Adversarial State Uncertainty in Deep Reinforcement Learning, *IEEE Transactions on Neural Networks and Learning Systems* (2021). DOI: 10.1109/TNNLS.2021.3056046

Provided by Massachusetts Institute of Technology

Citation: Algorithm helps artificial intelligence systems dodge 'adversarial' inputs (2021, March 8) retrieved 19 April 2024 from https://techxplore.com/news/2021-03-algorithm-artificial-intelligence-dodge-adversarial.html

This document is subject to copyright. Apart from any fair dealing for the purpose of private study or research, no part may be reproduced without the written permission. The content is provided for information purposes only.