Misinformation or artifact: A new way to think about machine learning

November 23 2020, by Jeannie Kever

Credit: CC0 Public Domain

Deep neural networks, multilayered systems built to process images and other data through the use of mathematical modeling, are a cornerstone of artificial intelligence.

They are capable of seemingly sophisticated results, but they can also be

fooled in ways that range from relatively harmless—misidentifying one animal as another—to potentially deadly if the network guiding a self-driving car misinterprets a stop sign as one indicating it is safe to proceed.

A philosopher with the University of Houston suggests in a paper published in *Nature Machine Intelligence* that common assumptions about the cause behind these supposed malfunctions may be mistaken, information that is crucial for evaluating the reliability of these networks.

As machine learning and other forms of artificial intelligence become more embedded in society, used in everything from automated teller machines to cybersecurity systems, Cameron Buckner, associate professor of philosophy at UH, said it is critical to understand the source of apparent failures caused by what researchers call "adversarial examples," when a deep neural network system misjudges images or other data when confronted with information outside the training inputs used to build the network. They're rare and are called "adversarial" because they are often created or discovered by another machine learning network—a sort of brinksmanship in the machine learning world between more sophisticated methods to create adversarial examples and more sophisticated methods to detect and avoid them.

"Some of these adversarial events could instead be artifacts, and we need to better know what they are in order to know how reliable these networks are," Buckner said.

In other words, the misfire could be caused by the interaction between what the network is asked to process and the actual patterns involved. That's not quite the same thing as being completely mistaken.

"Understanding the implications of adversarial examples requires

exploring a third possibility: that at least some of these patterns are artifacts," Buckner wrote. " ... Thus, there are presently both costs in simply discarding these patterns and dangers in using them naively."

Adversarial events that cause these machine learning systems to make mistakes aren't necessarily caused by intentional malfeasance, but that's where the highest risk comes in.

"It means malicious actors could fool systems that rely on an otherwise reliable network," Buckner said. "That has security applications."

A security system based upon facial recognition technology could be hacked to allow a breach, for example, or decals could be placed on traffic signs that cause self-driving cars to misinterpret the sign, even though they appear harmless to the human observer.

Previous research has found that, counter to previous assumptions, there are some naturally occurring adversarial examples—times when a machine learning system misinterprets data through an unanticipated interaction rather than through an error in the data. They are rare and can be discovered only through the use of artificial intelligence.

But they are real, and Buckner said that suggests the need to rethink how researchers approach the anomalies, or artifacts.

These artifacts haven't been well understood; Buckner offers the analogy of a lens flare in a photograph—a phenomenon that isn't caused by a defect in the camera lens but is instead produced by the interaction of light with the camera.

The lens flare potentially offers useful information—the location of the sun, for example—if you know how to interpret it. That, he said, raises the question of whether adverse events in machine learning that are

caused by an artifact also have useful information to offer.

Equally important, Buckner said, is that this new way of thinking about the way in which artifacts can affect deep neural networks suggests a misreading by the network shouldn't be automatically considered evidence that deep learning isn't valid.

"Some of these adversarial events could be artifacts," he said. "We have to know what these artifacts are so we can know how reliable the networks are."

More information: Buckner, C. Understanding adversarial examples requires a theory of artefacts for deep learning. *Nat Mach Intell* (2020). doi.org/10.1038/s42256-020-00266-y

Provided by University of Houston