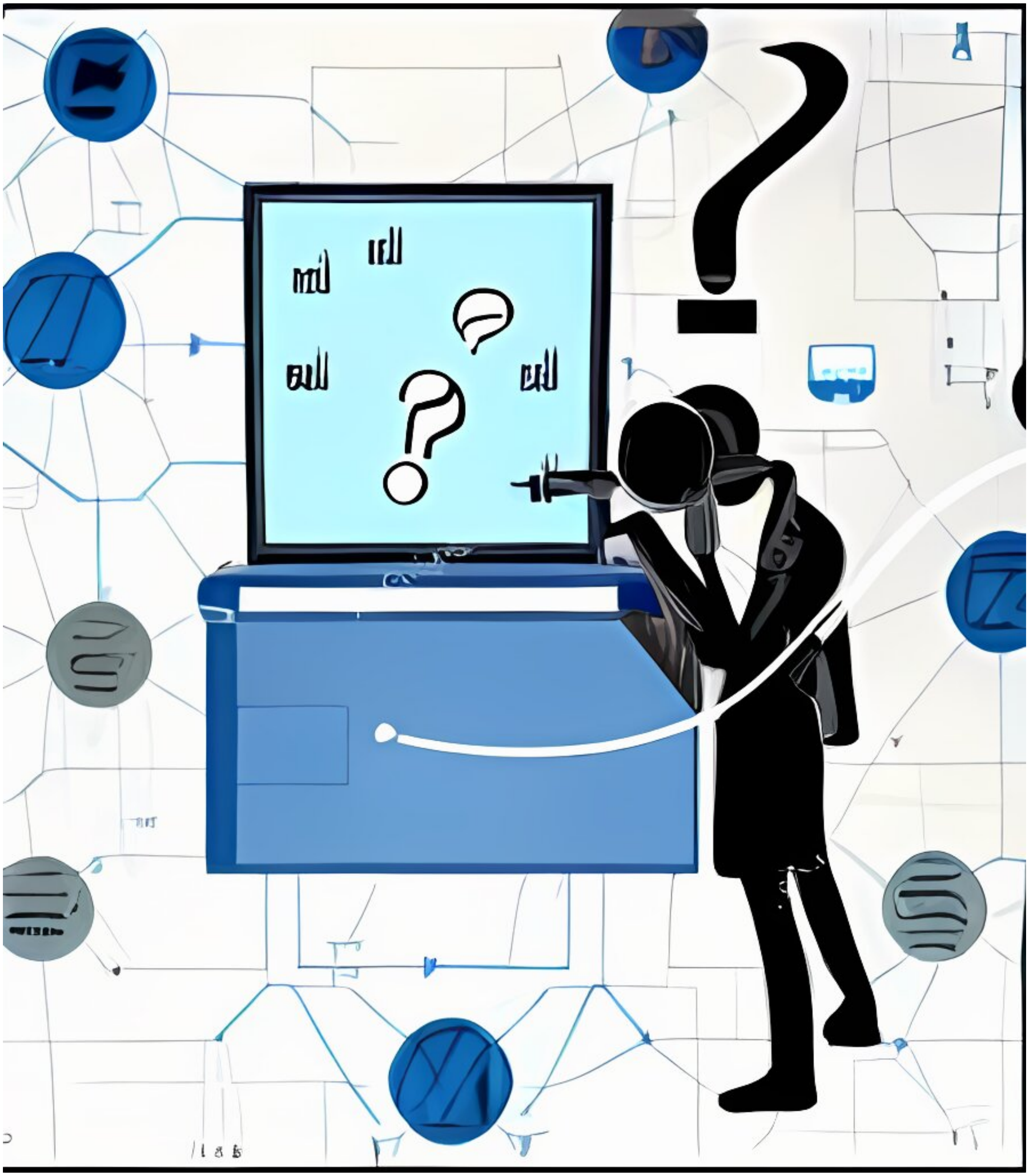
New method for comparing neural networks exposes how artificial intelligence works

September 13 2022



Researchers at Los Alamos are looking at new ways to compare neural networks. This image was created with an artificial intelligence software called Stable Diffusion, using the prompt "Peeking into the black box of neural networks." Credit: Los Alamos National Laboratory

A team at Los Alamos National Laboratory has developed a novel approach for comparing neural networks that looks within the "black box" of artificial intelligence to help researchers understand neural network behavior. Neural networks recognize patterns in datasets; they are used everywhere in society, in applications such as virtual assistants, facial recognition systems and self-driving cars.

"The artificial intelligence research community doesn't necessarily have a complete understanding of what neural networks are doing; they give us good results, but we don't know how or why," said Haydn Jones, a researcher in the Advanced Research in Cyber Systems group at Los Alamos. "Our new method does a better job of comparing neural networks, which is a crucial step toward better understanding the mathematics behind AI."

Jones is the lead author of the paper "If You've Trained One You've Trained Them All: Inter-Architecture Similarity Increases With Robustness," which was presented recently at the Conference on Uncertainty in Artificial Intelligence. In addition to studying network similarity, the paper is a crucial step toward characterizing the behavior of robust neural networks.

Neural networks are high-performance, but fragile. For example, self-driving cars use neural networks to detect signs. When conditions are ideal, they do this quite well. However, the smallest aberration—such as a sticker on a stop sign—can cause the neural network to misidentify the sign and never stop.

To improve neural networks, researchers are looking at ways to improve network robustness. One state-of-the-art approach involves "attacking" networks during their training process. Researchers intentionally introduce aberrations and train the AI to ignore them. This process is called adversarial training and essentially makes it harder to fool the networks.

Jones, Los Alamos collaborators Jacob Springer and Garrett Kenyon, and Jones' mentor Juston Moore, applied their new metric of network similarity to adversarially trained neural networks, and found, surprisingly, that adversarial training causes neural networks in the computer vision domain to converge to very similar data representations, regardless of network architecture, as the magnitude of the attack increases.

"We found that when we train neural networks to be robust against adversarial attacks, they begin to do the same things," Jones said.

There has been extensive effort in industry and in the academic community searching for the "right architecture" for neural networks, but the Los Alamos team's findings indicate that the introduction of adversarial training narrows this search space substantially. As a result, the AI research community may not need to spend as much time exploring new architectures, knowing that adversarial training causes diverse architectures to converge to similar solutions.

"By finding that robust neural networks are similar to each other, we're making it easier to understand how robust AI might really work. We might even be uncovering hints as to how perception occurs in humans and other animals," Jones said.

More information: Haydn T. Jones et al, If You've Trained One You've Trained Them All: Inter-Architecture Similarity Increases With Robustness, (2022)

Provided by Los Alamos National Laboratory

Citation: New method for comparing neural networks exposes how artificial intelligence works (2022, September 13) retrieved 8 May 2024 from https://techxplore.com/news/2022-09-method-neural-networks-exposes-artificial.html