Examining the mind's eye of a neural network system

November 16 2023

A diagnostic tool for neural networks makes finding errors as easy as spotting mountains from an airplane. Credit: Purdue University

In the background of image recognition software that can ID our friends on social media and wildflowers in our yard are neural networks, a type of artificial intelligence inspired by how own our brains process data.

While neural networks sprint through data, their architecture makes it

difficult to trace the origin of errors that are obvious to humans—like confusing a Converse high-top with an ankle boot—limiting their use in more vital work like health care image analysis or research. A new tool developed at Purdue University makes finding those errors as simple as spotting mountaintops from an airplane.

"In a sense, if a neural network were able to speak, we're showing you what it would be trying to say," said David Gleich, a Purdue professor of computer science in the College of Science who developed the tool, which is featured in a paper to published in *Nature Machine Intelligence*.

"The tool we've developed helps you find places where the network is saying, 'Hey, I need more information to do what you've asked.' I would advise people to use this tool on any high-stakes neural network decision scenarios or image prediction task."

Code for the tool is available on GitHub, as are use case demonstrations. Gleich collaborated on the research with Tamal K. Dey, also a Purdue professor of computer science, and Meng Liu, a former Purdue graduate student who earned a doctorate in computer science.

In testing their approach, Gleich's team caught neural networks mistaking the identity of images in databases of everything from chest X-rays and gene sequences to apparel. In one example, a neural network repeatedly mislabeled images of cars from the Imagenette database as cassette players. The reason? The pictures were drawn from online sales listings and included tags for the cars' stereo equipment.

Neural network image recognition systems are essentially algorithms that process data in a way that mimics the weighted firing pattern of neurons as an image is analyzed and identified. A system is trained to its task—such as identifying an animal, a garment or a tumor—with a "training set" of images that includes data on each pixel, tagging and

other information, and the identity of the image as classified within a particular category.

Using the training set, the network learns, or "extracts," the information it needs in order to match the input values with the category. This information, a string of numbers called an embedded vector, is used to calculate the probability that the image belongs to each of the possible categories. Generally speaking, the correct identity of the image is within the category with the highest probability.

But the embedded vectors and probabilities don't correlate to a decision-making process that humans would recognize. Feed in 100,000 numbers representing the known data, and the network produces an embedded vector of 128 numbers that don't correspond to physical features, although they do make it possible for the network to classify the image.

In other words, you can't open the hood on the algorithms of a trained system and follow along. Between the input values and the predicted identity of the image is a proverbial "black box" of unrecognizable numbers across multiple layers.

"The problem with neural networks is that we can't see inside the machine to understand how it's making decisions, so how can we know if a neural network is making a characteristic mistake?" Gleich said.

Rather than trying to trace the decision-making path of any single image through the network, Gleich's approach makes it possible to visualize the relationship that the computer sees among all the images in an entire database. Think of it like a bird's-eye view of all the images as the neural network has organized them.

The relationship among the images (like network's prediction of the identity classification of each of the images in the database) is based on

the embedded vectors and probabilities the network generates. To boost the resolution of the view and find places where the network can't distinguish between two different classifications, Gleich's team first developed a method of splitting and overlapping the classifications to identify where images have a high probability of belonging to more than one classification.

The team then maps the relationships onto a Reeb graph, a tool taken from the field of topological data analysis. On the graph, each group of images the network thinks are related is represented by a single dot. Dots are color coded by classification. The closer the dots, the more similar the network considers groups to be, and most areas of the graph show clusters of dots in a single color.

But groups of images with a high probability of belonging to more than one classification will be represented by two differently colored overlapping dots. With a single glance, areas where the network cannot distinguish between two classifications appear as a cluster of dots in one color, accompanied by a smattering of overlapping dots in a second color. Zooming in on the overlapping dots will show an area of confusion, like the picture of the car that's been labeled both car and cassette player.

"What we're doing is taking these complicated sets of information coming out of the network and giving people an 'in' into how the network sees the data at a macroscopic level," Gleich said. "The Reeb map represents the important things, the big groups and how they relate to each other, and that makes it possible to see the errors."

More information: Meng Liu et al, Topological structure of complex predictions, *Nature Machine Intelligence* (2023). DOI: 10.1038/s42256-023-00749-8. On *arXiv*: DOI: 10.48550/arxiv.2207.14358

Provided by Purdue University