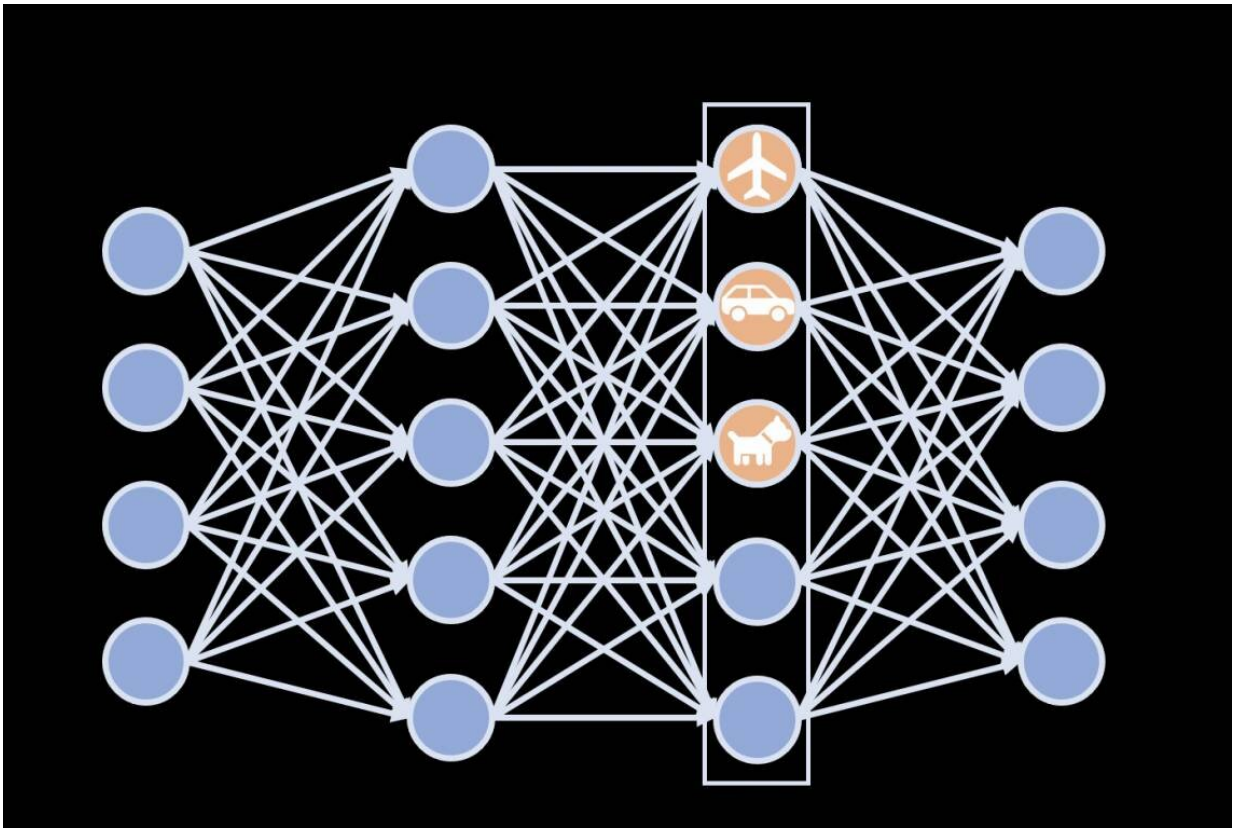
Accurate neural network computer vision without the 'black box'

December 15 2020



New research offers clues to what goes on inside the minds of machines as they learn to see. A method developed by Cynthia Rudin's lab reveals how much a neural network calls to mind different concepts as an image travels through the network's layers. Credit: Duke University School of Nursing

The artificial intelligence behind self-driving cars, medical image

analysis and other computer vision applications relies on what's called deep neural networks.

Loosely modeled on the brain, these consist of layers of interconnected "neurons"—mathematical functions that send and receive information—that "fire" in response to features of the input data. The first layer processes a raw data input—such as pixels in an image—and passes that information to the next layer above, triggering some of those neurons, which then pass a signal to even higher layers until eventually it arrives at a determination of what is in the input image.

But here's the problem, says Duke computer science professor Cynthia Rudin. "We can input, say, a medical image, and observe what comes out the other end ('this is a picture of a malignant lesion', but it's hard to know what happened in between."

It's what's known as the "black box" problem. What happens in the mind of the machine—the network's hidden layers—is often inscrutable, even to the people who built it.

"The problem with deep learning models is they're so complex that we don't actually know what they're learning," said Zhi Chen, a Ph.D. student in Rudin's lab at Duke. "They can often leverage information we don't want them to. Their reasoning processes can be completely wrong."

Rudin, Chen and Duke undergraduate Yijie Bei have come up with a way to address this issue. By modifying the reasoning process behind the predictions, it is possible that researchers can better troubleshoot the networks or understand whether they are trustworthy.

Most approaches attempt to uncover what led a computer vision system to the right answer after the fact, by pointing to the key features or pixels that identified an image: "The growth in this chest X-ray was

classified as malignant because, to the model, these areas are critical in the classification of lung cancer." Such approaches don't reveal the network's reasoning, just where it was looking.

The Duke team tried a different tack. Instead of attempting to account for a network's decision-making on a post hoc basis, their method trains the network to show its work by expressing its understanding about concepts along the way. Their method works by revealing how much the network calls to mind different concepts to help decipher what it sees. "It disentangles how different concepts are represented within the layers of the network," Rudin said.

Given an image of a library, for example, the approach makes it possible to determine whether and how much the different layers of the neural network rely on their mental representation of "books" to identify the scene.

The researchers found that, with a small adjustment to a neural network, it is possible to identify objects and scenes in images just as accurately as the original network, and yet gain substantial interpretability in the network's reasoning process. "The technique is very simple to apply," Rudin said.

The method controls the way information flows through the network. It involves replacing one standard part of a neural network with a new part. The new part constrains only a single neuron in the network to fire in response to a particular concept that humans understand. The concepts could be categories of everyday objects, such as "book" or "bike." But they could also be general characteristics, such as such as "metal," "wood," "cold" or "warm." By having only one neuron control the information about one concept at a time, it is much easier to understand how the network "thinks."

The researchers tried their approach on a neural network trained by millions of labeled images to recognize various kinds of indoor and outdoor scenes, from classrooms and food courts to playgrounds and patios. Then they turned it on images it hadn't seen before. They also looked to see which concepts the network layers drew on the most as they processed the data.

Chen pulls up a plot showing what happened when they fed a picture of an orange sunset into the network. Their trained neural network says that warm colors in the sunset image, like orange, tend to be associated with the concept "bed" in earlier layers of the network. In short, the network activates the "bed neuron" highly in early layers. As the image travels through successive layers, the network gradually relies on a more sophisticated mental representation of each concept, and the "airplane" concept becomes more activated than the notion of beds, perhaps because "airplanes" are more often associated with skies and clouds.

It's only a small part of what's going on, to be sure. But from this trajectory the researchers are able to capture important aspects of the network's train of thought.

The researchers say their module can be wired into any neural network that recognizes images. In one experiment, they connected it to a neural network trained to detect skin cancer in photos.

Before an AI can learn to spot melanoma, it must learn what makes melanomas look different from normal moles and other benign spots on your skin, by sifting through thousands of training images labeled and marked up by skin cancer experts.

But the network appeared to be summoning up a concept of "irregular border" that it formed on its own, without help from the training labels. The people annotating the images for use in artificial intelligence

applications hadn't made note of that feature, but the machine did.

"Our method revealed a shortcoming in the dataset," Rudin said. Perhaps if they had included this information in the data, it would have made it clearer whether the model was reasoning correctly. "This example just illustrates why we shouldn't put blind faith in "black box" models with no clue of what goes on inside them, especially for tricky medical diagnoses," Rudin said.

The team's work appeared Dec. 7 in the journal *Nature Machine Intelligence*.

More information: Zhi Chen et al, Concept whitening for interpretable image recognition, *Nature Machine Intelligence* (2020). DOI: 10.1038/s42256-020-00265-z

Provided by Duke University School of Nursing