New technique elucidates the inner workings of neural networks trained on visual data

June 30 2017, by Larry Hardesty

Neural networks learn to perform computational tasks by analyzing large sets of training data. But once they've been trained, even their designers rarely have any idea what data elements they're processing. Credit: Christine Daniloff/MIT

Neural networks, which learn to perform computational tasks by

analyzing large sets of training data, are responsible for today's best-performing artificial intelligence systems, from speech recognition systems, to automatic translators, to self-driving cars.

But neural nets are black boxes. Once they've been trained, even their designers rarely have any idea what they're doing—what data elements they're processing and how.

Two years ago, a team of computer-vision researchers from MIT's Computer Science and Artificial Intelligence Laboratory (CSAIL) described a method for peering into the black box of a neural net trained to identify visual scenes. The method provided some interesting insights, but it required data to be sent to human reviewers recruited through Amazon's Mechanical Turk crowdsourcing service.

At this year's Computer Vision and Pattern Recognition conference, CSAIL researchers will present a fully automated version of the same system. Where the previous paper reported the analysis of one type of neural network trained to perform one task, the new paper reports the analysis of four types of neural networks trained to perform more than 20 tasks, including recognizing scenes and objects, colorizing grey images, and solving puzzles. Some of the new networks are so large that analyzing any one of them would have been cost-prohibitive under the old method.

The researchers also conducted several sets of experiments on their networks that not only shed light on the nature of several computer-vision and computational-photography algorithms, but could also provide some evidence about the organization of the human brain.

Neural networks are so called because they loosely resemble the human nervous system, with large numbers of fairly simple but densely connected information-processing "nodes." Like neurons, a neural net's

nodes receive information signals from their neighbors and then either "fire"—emitting their own signals—or don't. And as with neurons, the strength of a node's firing response can vary.

In both the new paper and the earlier one, the MIT researchers doctored neural networks trained to perform computer vision tasks so that they disclosed the strength with which individual nodes fired in response to different input images. Then they selected the 10 input images that provoked the strongest response from each node.

In the earlier paper, the researchers sent the images to workers recruited through Mechanical Turk, who were asked to identify what the images had in common. In the new paper, they use a computer system instead.

"We catalogued 1,100 visual concepts—things like the color green, or a swirly texture, or wood material, or a human face, or a bicycle wheel, or a snowy mountaintop," says David Bau, an MIT graduate student in electrical engineering and computer science and one of the paper's two first authors. "We drew on several data sets that other people had developed, and merged them into a broadly and densely labeled data set of visual concepts. It's got many, many labels, and for each label we know which pixels in which image correspond to that label."

The paper's other authors are Bolei Zhou, co-first author and fellow graduate student; Antonio Torralba, MIT professor of electrical engineering and computer science; Aude Oliva, CSAIL principal research scientist; and Aditya Khosla, who earned his PhD as a member of Torralba's group and is now the chief technology officer of the medical-computing company PathAI.

The researchers also knew which pixels of which images corresponded to a given network node's strongest responses. Today's neural nets are organized into layers. Data are fed into the lowest layer, which processes

them and passes them to the next layer, and so on. With visual data, the input images are broken into small chunks, and each chunk is fed to a separate input node.

For every strong response from a high-level node in one of their networks, the researchers could trace back the firing patterns that led to it, and thus identify the specific image pixels it was responding to. Because their system could frequently identify labels that corresponded to the precise pixel clusters that provoked a strong response from a given node, it could characterize the node's behavior with great specificity.

The researchers organized the visual concepts in their database into a hierarchy. Each level of the hierarchy incorporates concepts from the level below, beginning with colors and working upward through textures, materials, parts, objects, and scenes. Typically, lower layers of a neural network would fire in response to simpler visual properties—such as colors and textures—and higher layers would fire in response to more complex properties.

But the hierarchy also allowed the researchers to quantify the emphasis that networks trained to perform different tasks placed on different visual properties. For instance, a network trained to colorize black-and-white images devoted a large majority of its nodes to recognizing textures. Another network, when trained to track objects across several frames of video, devoted a higher percentage of its nodes to scene recognition than it did when trained to recognize scenes; in that case, many of its nodes were in fact dedicated to object detection.

One of the researchers' experiments could conceivably shed light on a vexed question in neuroscience. Research involving human subjects with electrodes implanted in their brains to control severe neurological disorders has seemed to suggest that individual neurons in the brain fire in response to specific visual stimuli. This hypothesis, originally called

the grandmother-neuron hypothesis, is more familiar to a recent generation of neuroscientists as the Jennifer-Aniston-neuron hypothesis, after the discovery that several neurological patients had neurons that appeared to respond only to depictions of particular Hollywood celebrities.

Many neuroscientists dispute this interpretation. They argue that shifting constellations of neurons, rather than individual neurons, anchor sensory discriminations in the brain. Thus, the so-called Jennifer Aniston neuron is merely one of many neurons that collectively fire in response to images of Jennifer Aniston. And it's probably part of many other constellations that fire in response to stimuli that haven't been tested yet.

Because their new analytic technique is fully automated, the MIT researchers were able to test whether something similar takes place in a neural network trained to recognize visual scenes. In addition to identifying individual network nodes that were tuned to particular visual concepts, they also considered randomly selected combinations of nodes. Combinations of nodes, however, picked out far fewer visual concepts than individual nodes did—roughly 80 percent fewer.

"To my eye, this is suggesting that neural networks are actually trying to approximate getting a grandmother neuron," Bau says. "They're not trying to just smear the idea of grandmother all over the place. They're trying to assign it to a neuron. It's this interesting hint of this structure that most people don't believe is that simple."

More information: Network Dissection: Quantifying Interpretability of Deep Visual Representations. netdissect.csail.mit.edu/final … twork-dissection.pdf

This story is republished courtesy of MIT News

(web.mit.edu/newsoffice/), a popular site that covers news about MIT research, innovation and teaching.

Provided by Massachusetts Institute of Technology