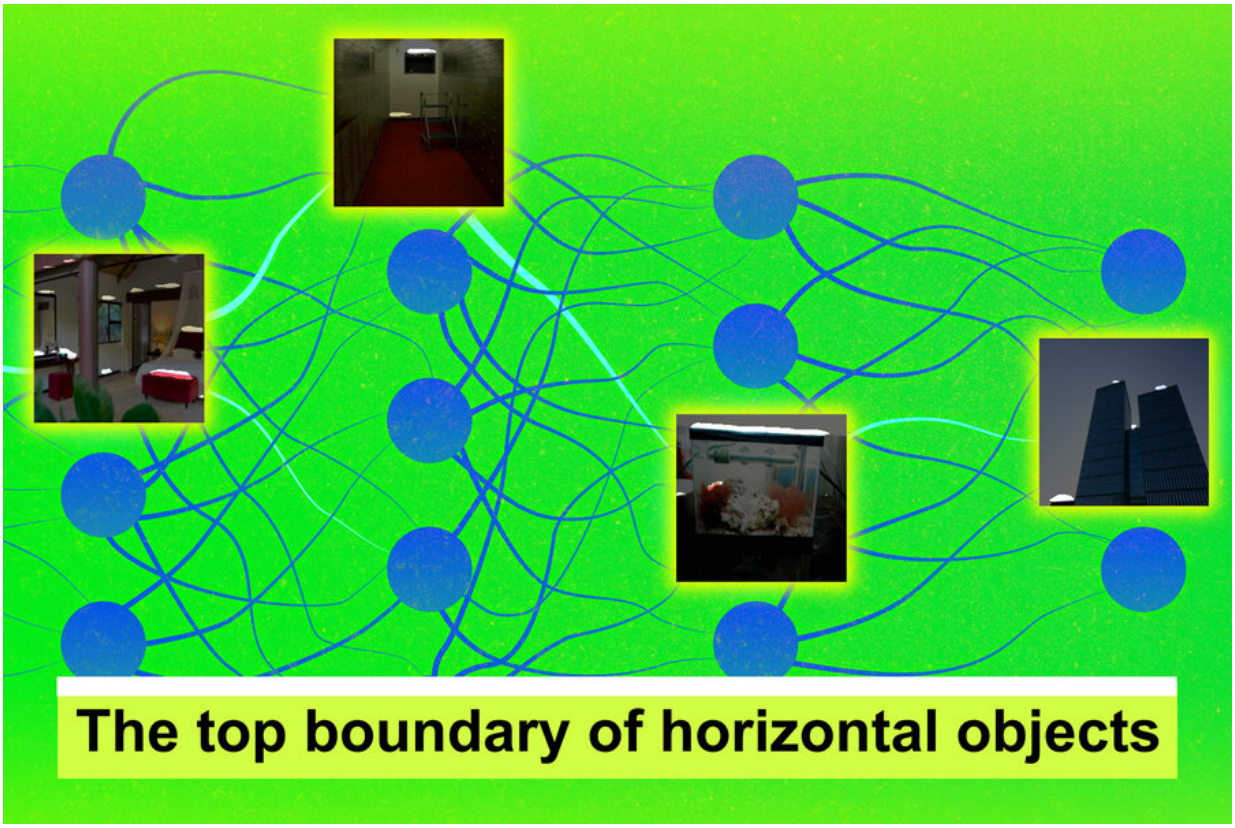
Demystifying machine-learning systems using natural language

January 27 2022, by Adam Zewe



MIT researchers created a technique that can automatically describe the roles of individual neurons in a neural network with natural language. In this figure, the technique was able to identify “the top boundary of horizontal objects” in photographs, which are highlighted in white. Credit: Jose-Luis Olivares, MIT

Neural networks are sometimes called black boxes because, despite the

fact that they can outperform humans on certain tasks, even the researchers who design them often don't understand how or why they work so well. But if a neural network is used outside the lab, perhaps to classify medical images that could help diagnose heart conditions, knowing how the model works helps researchers predict how it will behave in practice.

MIT researchers have now developed a method that sheds some light on the inner workings of black box neural networks. Modeled off the human brain, neural networks are arranged into layers of interconnected nodes, or "neurons," that process data. The new system can automatically produce descriptions of those individual neurons, generated in English or another natural language.

For instance, in a neural network trained to recognize animals in images, their method might describe a certain neuron as detecting ears of foxes. Their scalable technique is able to generate more accurate and specific descriptions for individual neurons than other methods.

In a new paper, the team shows that this method can be used to audit a neural network to determine what it has learned, or even edit a network by identifying and then switching off unhelpful or incorrect neurons.

"We wanted to create a method where a machine-learning practitioner can give this system their model and it will tell them everything it knows about that model, from the perspective of the model's neurons, in language. This helps you answer the basic question, 'Is there something my model knows about that I would not have expected it to know?'" says Evan Hernandez, a graduate student in the MIT Computer Science and Artificial Intelligence Laboratory (CSAIL) and lead author of the paper.

Automatically generated descriptions

Most existing techniques that help machine-learning practitioners understand how a model works either describe the entire neural network or require researchers to identify concepts they think individual neurons could be focusing on.

The system Hernandez and his collaborators developed, dubbed MILAN (mutual-information guided linguistic annotation of neurons), improves upon these methods because it does not require a list of concepts in advance and can automatically generate natural language descriptions of all the neurons in a network. This is especially important because one neural network can contain hundreds of thousands of individual neurons.

MILAN produces descriptions of neurons in neural networks trained for computer vision tasks like object recognition and image synthesis. To describe a given neuron, the system first inspects that neuron's behavior on thousands of images to find the set of image regions in which the neuron is most active. Next, it selects a natural language description for each neuron to maximize a quantity called pointwise mutual information between the image regions and descriptions. This encourages descriptions that capture each neuron's distinctive role within the larger network.

"In a neural network that is trained to classify images, there are going to be tons of different neurons that detect dogs. But there are lots of different types of dogs and lots of different parts of dogs. So even though 'dog' might be an accurate description of a lot of these neurons, it is not very informative. We want descriptions that are very specific to what that neuron is doing. This isn't just dogs; this is the left side of ears on German shepherds," says Hernandez.

The team compared MILAN to other models and found that it generated richer and more accurate descriptions, but the researchers were more

interested in seeing how it could assist in answering specific questions about computer vision models.

Analyzing, auditing, and editing neural networks

First, they used MILAN to analyze which neurons are most important in a neural network. They generated descriptions for every neuron and sorted them based on the words in the descriptions. They slowly removed neurons from the network to see how its accuracy changed, and found that neurons that had two very different words in their descriptions (vases and fossils, for instance) were less important to the network.

They also used MILAN to audit models to see if they learned something unexpected. The researchers took image classification models that were trained on datasets in which human faces were blurred out, ran MILAN, and counted how many neurons were nonetheless sensitive to human faces.

"Blurring the faces in this way does reduce the number of neurons that are sensitive to faces, but far from eliminates them. As a matter of fact, we hypothesize that some of these face neurons are very sensitive to specific demographic groups, which is quite surprising. These models have never seen a human face before, and yet all kinds of facial processing happens inside them," Hernandez says.

In a third experiment, the team used MILAN to edit a neural network by finding and removing neurons that were detecting bad correlations in the data, which led to a 5 percent increase in the network's accuracy on inputs exhibiting the problematic correlation.

While the researchers were impressed by how well MILAN performed in these three applications, the model sometimes gives descriptions that

are still too vague, or it will make an incorrect guess when it doesn't know the concept it is supposed to identify.

They are planning to address these limitations in future work. They also want to continue enhancing the richness of the descriptions MILAN is able to generate. They hope to apply MILAN to other types of neural networks and use it to describe what groups of neurons do, since neurons work together to produce an output.

"This is an approach to interpretability that starts from the bottom up. The goal is to generate open-ended, compositional descriptions of function with natural language. We want to tap into the expressive power of human language to generate descriptions that are a lot more natural and rich for what neurons do. Being able to generalize this approach to different types of models is what I am most excited about," says Schwettmann.

"The ultimate test of any technique for explainable AI is whether it can help researchers and users make better decisions about when and how to deploy AI systems," says Andreas. "We're still a long way off from being able to do that in a general way. But I'm optimistic that MILAN—and the use of language as an explanatory tool more broadly—will be a useful part of the toolbox."

The research was published on *arXiv*.

More information: Evan Hernandez et al, Natural Language Descriptions of Deep Visual Features, arXiv:2201.11114 [cs.CV], arxiv.org/abs/2201.11114

research, innovation and teaching.

Provided by Massachusetts Institute of Technology