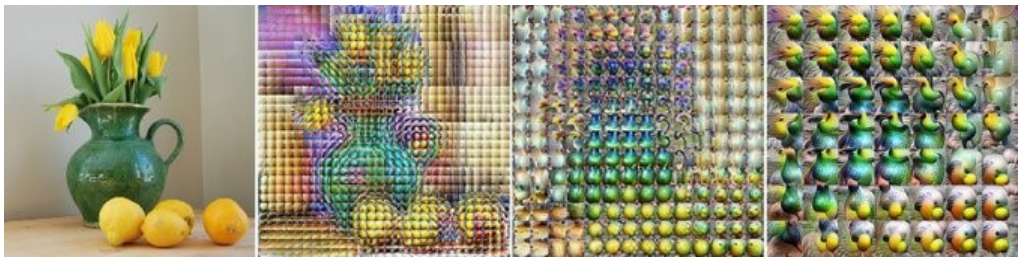
Google Researchers Are Learning How Machines Learn

Cade Metz

On the left is an image that was put through a neural network trained to classify objects in images — for example, to tell whether an image includes a vase or a lemon. On the right is a visualization of what one layer in the middle of the network detected at each position of the image. The neural network seems to be detecting vase-like patterns and lemon-like objects. The Building Blocks of Interpretability

SAN FRANCISCO — Machines are starting to learn tasks on their own. They are identifying faces, recognizing spoken words, reading medical scans and even carrying on their own conversations.



The original image and three more visualizations after it was put through a neural network. The first layer is primarily detecting edges and color. The other layers begin recognizing more complex concepts like flowers, vases and lemons. The Building Blocks of Interpretability

All this is done through so-called neural networks, which are complex computer algorithms that learn tasks by analyzing vast amounts of data. But these neural networks create a problem that scientists are trying to

solve: It is not always easy to tell how the machines arrive at their conclusions.

On Tuesday, a team at Google took a small step toward addressing this issue with the unveiling of new research that offers the rough outlines of technology that shows how the machines are arriving at their decisions.

Groups of neurons automatically learn to work together to represent concepts in an image. Five groups of neurons seem to correspond to flowers, the lip of the vase, the body of vase, the background, and lemons. A heat map shows where each neuron group fired on the image. The Building Blocks of Interpretability

“Even seeing part of how a decision was made can give you a lot of insight into the possible ways it can fail,” said Christopher Olah, a Google researcher.

A growing number of A.I. researchers are now developing ways to better understand neural networks. Jeff Clune, a professor at University of Wyoming who now works in the A.I. lab at the ride-hailing company Uber, called this “artificial neuroscience.”

Neuron groups at two different layers of the network and the output classes. The lines show which neuron groups support or inhibit later groups or output classes. For example, a \"lemon\" classification is strongly supported by a yellow, lemon-y group. The Building Blocks of Interpretability

Understanding how these systems work will become more important as they make decisions now made by humans, like who gets a job and how a self-driving car responds to emergencies.

First proposed in the 1950s, neural networks are meant to mimic the web of neurons in the brain. But that is a rough analogy. These algorithms are really series of mathematical operations, and each operation represents a neuron. Google's new research aims to show — in a highly visual way — how these mathematical operations perform discrete tasks, like recognizing objects in photos.

A \"vase\" classification is supported by the groups that represent flowers, the lip of the vase and the background. The Building Blocks of Interpretability

Inside a neural network, each neuron works to identify a particular characteristic that might show up in a photo, like a line that curves from right to left at a certain angle or several lines that merge to form a larger shape. Google wants to provide tools that show what each neuron is trying to identify, which ones are successful and how their efforts combine to determine what is actually in the photo — perhaps a dog or a tuxedo or a bird.

The kind of technology Google is discussing could also help identify why a neural network is prone to mistakes and, in some cases, explain how it learned this behavior, Mr. Olah said. Other researchers, including Mr. Clune, believe they can also help minimize the threat of “adversarial examples” — where someone can potentially fool neural networks by, say, doctoring an image.

Researchers acknowledge that this work is still in its infancy. Jason Yosinski, who also works in Uber’s A.I. lab, which grew out of the company’s acquisition of a start-up called Geometric Intelligence, called Google’s technology idea “state of art.” But he warned it may never be entirely easy to understand the computer mind.

“To a certain extent, as these networks get more complicated, it is going to be fundamentally difficult to understand why they make decisions,” he said. “It is kind of like trying to understand why humans make decisions.”