Neuroscientists find a way to make object-recognition models perform better

December 3 2020, by Anne Trafton

MIT neuroscientists have developed a way to overcome computer vision models' vulnerability to "adversarial attacks," by adding to these models a new layer that is designed to mimic V1, the earliest stage of the brain's visual processing system. Credit: MIT News.

Computer vision models known as convolutional neural networks can be

trained to recognize objects nearly as accurately as humans do. However, these models have one significant flaw: Very small changes to an image, which would be nearly imperceptible to a human viewer, can trick them into making egregious errors such as classifying a cat as a tree.

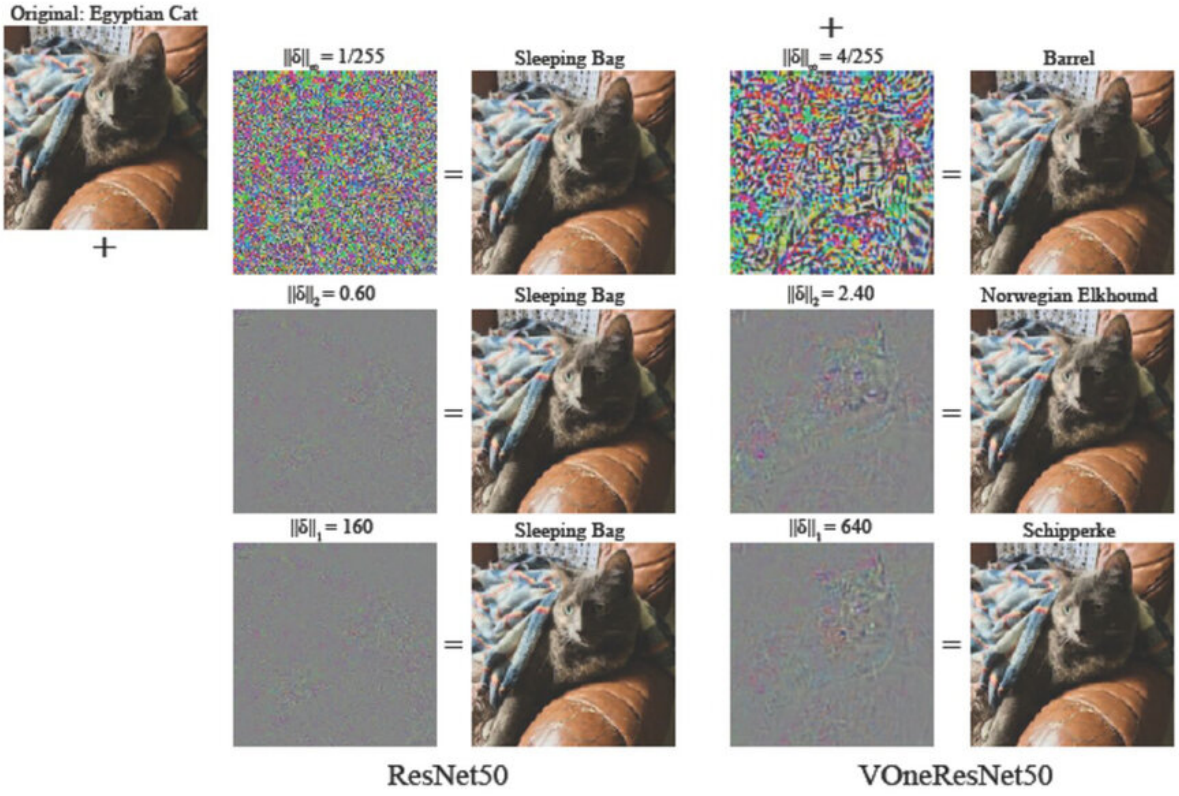
A team of neuroscientists from MIT, Harvard University, and IBM have developed a way to alleviate this vulnerability, by adding to these models a new layer that is designed to mimic the earliest stage of the brain's visual processing system. In a new study, they showed that this layer greatly improved the models' robustness against this type of mistake.

"Just by making the models more similar to the brain's primary visual cortex, in this single stage of processing, we see quite significant improvements in robustness across many different types of perturbations and corruptions," says Tiago Marques, an MIT postdoc and one of the lead authors of the study.

Convolutional neural networks are often used in artificial intelligence applications such as self-driving cars, automated assembly lines, and medical diagnostics. Harvard graduate student Joel Dapello, who is also a lead author of the study, adds that "implementing our new approach could potentially make these systems less prone to error and more aligned with human vision."

"Good scientific hypotheses of how the brain's visual system works should, by definition, match the brain in both its internal neural patterns and its remarkable robustness. This study shows that achieving those scientific gains directly leads to engineering and application gains," says James DiCarlo, the head of MIT's Department of Brain and Cognitive Sciences, an investigator in the Center for Brains, Minds, and Machines and the McGovern Institute for Brain Research, and the senior author of the study.

The study, which is being presented at the NeurIPS conference this month, is also co-authored by MIT graduate student Martin Schrimpf, MIT visiting student Franziska Geiger, and MIT-IBM Watson AI Lab Director David Cox.



A comparison of adversarial images with different perturbation strengths. Credit: Massachusetts Institute of Technology

Mimicking the brain

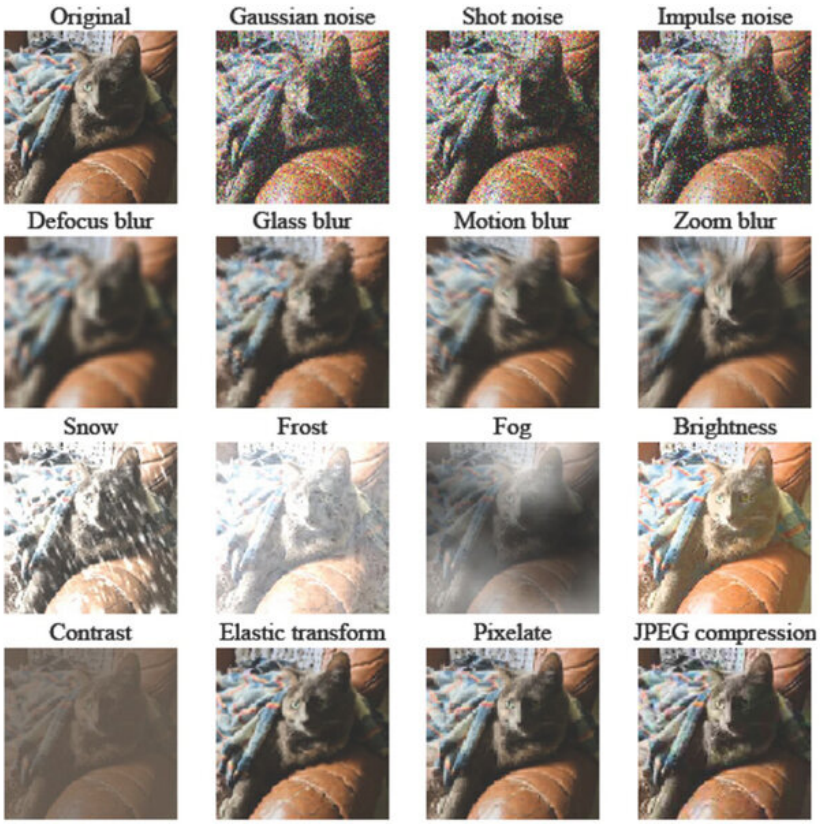
Recognizing objects is one of the visual system's primary functions. In just a small fraction of a second, visual information flows through the ventral visual stream to the brain's inferior temporal cortex, where

neurons contain information needed to classify objects. At each stage in the ventral stream, the brain performs different types of processing. The very first stage in the ventral stream, V1, is one of the most well-characterized parts of the brain and contains neurons that respond to simple visual features such as edges.

"It's thought that V1 detects local edges or contours of objects, and textures, and does some type of segmentation of the images at a very small scale. Then that information is later used to identify the shape and texture of objects downstream," Marques says. "The visual system is built in this hierarchical way, where in early stages neurons respond to local features such as small, elongated edges."

For many years, researchers have been trying to build computer models that can identify objects as well as the human visual system. Today's leading computer vision systems are already loosely guided by our current knowledge of the brain's visual processing. However, neuroscientists still don't know enough about how the entire ventral visual stream is connected to build a model that precisely mimics it, so they borrow techniques from the field of machine learning to train convolutional neural networks on a specific set of tasks. Using this process, a model can learn to identify objects after being trained on millions of images.

Many of these convolutional networks perform very well, but in most cases, researchers don't know exactly how the network is solving the object-recognition task. In 2013, researchers from DiCarlo's lab showed that some of these neural networks could not only accurately identify objects, but they could also predict how neurons in the primate brain would respond to the same objects much better than existing alternative models. However, these neural networks are still not able to perfectly predict responses along the ventral visual stream, particularly at the earliest stages of object recognition, such as V1.



A grid showing the visualization of many common image corruption types. First row, original image, followed by the noise corruptions; second row, blur corruptions; third row, weather corruptions; fourth row, digital corruptions. Credit: Massachusetts Institute of Technology

These models are also vulnerable to so-called "adversarial attacks." This means that small changes to an image, such as changing the colors of a few pixels, can lead the model to completely confuse an object for something different—a type of mistake that a human viewer would not make.

As a first step in their study, the researchers analyzed the performance of 30 of these models and found that models whose internal responses

better matched the brain's V1 responses were also less vulnerable to adversarial attacks. That is, having a more brain-like V1 seemed to make the model more robust. To further test and take advantage of that idea, the researchers decided to create their own model of V1, based on existing neuroscientific models, and place it at the front of convolutional neural networks that had already been developed to perform object recognition.

When the researchers added their V1 layer, which is also implemented as a convolutional neural network, to three of these models, they found that these models became about four times more resistant to making mistakes on images perturbed by adversarial attacks. The models were also less vulnerable to misidentifying objects that were blurred or distorted due to other corruptions.

"Adversarial attacks are a big, open problem for the practical deployment of deep neural networks. The fact that adding neuroscience-inspired elements can improve robustness substantially suggests that there is still a lot that AI can learn from neuroscience, and vice versa," Cox says.

More information: Simulating a Primary Visual Cortex at the Front of CNNs Improves Robustness to Image Perturbations. proceedings.neurips.cc/paper/2 … 470841-Abstract.html

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