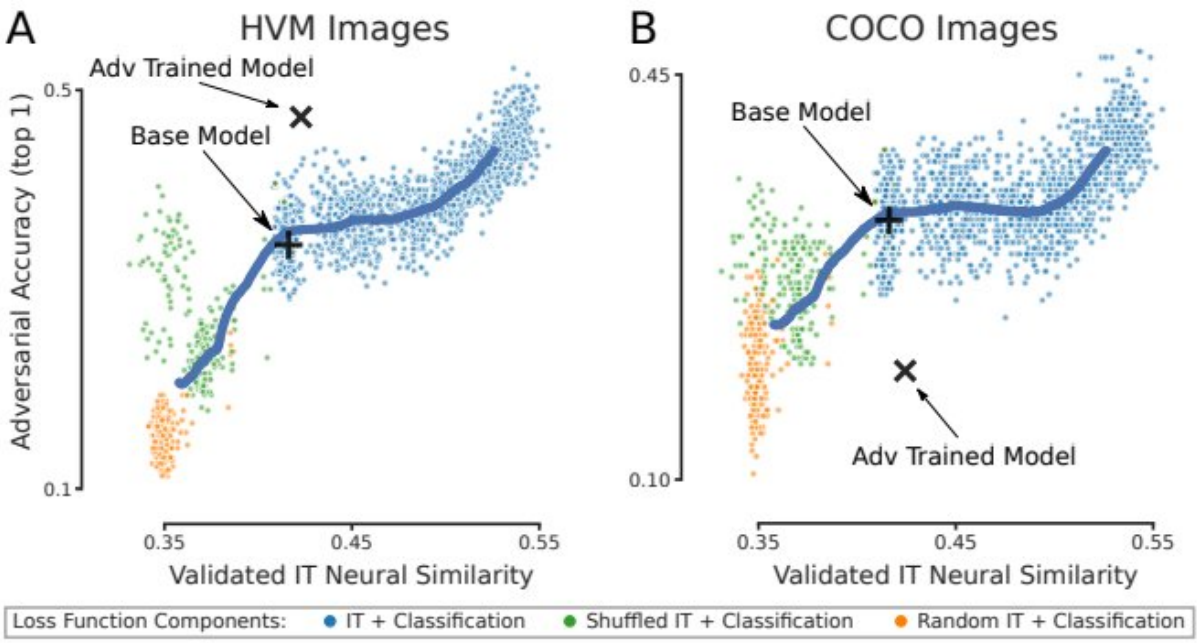
When computer vision works more like a brain, it sees more like people do

July 6 2023, by Jennifer Michalowski



IT neural similarity correlates with improved white box adversarial robustness. A) held out animal and image IT neural similarity is plotted against white box adversarial accuracy (PGD L_∞ ϵ = 1/1020) on the HVM image set measured across multiple training time points for all neural loss ratio conditions, random Gaussian IT target matrix conditions, and image shuffled IT target matrix conditions. B) Like in A but for COCO images. In both plots, the black cross represents the average base model position, the black X marks a CORnet-S adversarially trained on HVM images, and the heavy blue line is a sliding X, Y average of all conditions merely to visually highlight trends. Five seeds for each condition are plotted. Credit: Aligning Model and Macaque Inferior Temporal Cortex Representations Improves Model-to-Human Behavioral Alignment and

Adversarial Robustness.
https://openreview.net/attachment?id=SMYdcXjJh1q&name=pdf

From cameras to self-driving cars, many of today's technologies depend on artificial intelligence to extract meaning from visual information. Today's AI technology has artificial neural networks at its core, and most of the time we can trust these AI computer vision systems to see things the way we do—but sometimes they falter. According to MIT and IBM research scientists, one way to improve computer vision is to instruct the artificial neural networks that they rely on to deliberately mimic the way the brain's biological neural network processes visual images.

Researchers led by MIT Professor James DiCarlo, the director of MIT's Quest for Intelligence and member of the MIT-IBM Watson AI Lab, have made a computer vision model more robust by training it to work like a part of the brain that humans and other primates rely on for object recognition. This May, at the International Conference on Learning Representations, the team reported that when they trained an artificial neural network using neural activity patterns in the brain's inferior temporal (IT) cortex, the artificial neural network was more robustly able to identify objects in images than a model that lacked that neural training. And the model's interpretations of images more closely matched what humans saw, even when images included minor distortions that made the task more difficult.

Comparing neural circuits

Many of the artificial neural networks used for computer vision already resemble the multilayered brain circuits that process visual information in humans and other primates. Like the brain, they use neuron-like units that work together to process information. As they are trained for a

particular task, these layered components collectively and progressively process the visual information to complete the task—determining, for example, that an image depicts a bear or a car or a tree.

DiCarlo and others previously found that when such deep-learning computer vision systems establish efficient ways to solve visual problems, they end up with artificial circuits that work similarly to the neural circuits that process visual information in our own brains. That is, they turn out to be surprisingly good scientific models of the neural mechanisms underlying primate and human vision.

That resemblance is helping neuroscientists deepen their understanding of the brain. By demonstrating ways visual information can be processed to make sense of images, computational models suggest hypotheses about how the brain might accomplish the same task. As developers continue to refine computer vision models, neuroscientists have found new ideas to explore in their own work.

"As vision systems get better at performing in the real world, some of them turn out to be more human-like in their internal processing. That's useful from an understanding-biology point of view," says DiCarlo, who is also a professor of brain and cognitive sciences and an investigator at the McGovern Institute for Brain Research.

Engineering a more brain-like AI

While their potential is promising, computer vision systems are not yet perfect models of human vision. DiCarlo suspected one way to improve computer vision may be to incorporate specific brain-like features into these models.

To test this idea, he and his collaborators built a computer vision model using neural data previously collected from vision-processing neurons in

the monkey IT cortex—a key part of the primate ventral visual pathway involved in the recognition of objects—while the animals viewed various images. More specifically, Joel Dapello, a Harvard University graduate student and former MIT-IBM Watson AI Lab intern; and Kohitij Kar, assistant professor and Canada Research Chair (Visual Neuroscience) at York University and visiting scientist at MIT; in collaboration with David Cox, IBM Research's vice president for AI models and IBM director of the MIT-IBM Watson AI Lab; and other researchers at IBM Research and MIT asked an artificial neural network to emulate the behavior of these primate vision-processing neurons while the network learned to identify objects in a standard computer vision task.

"In effect, we said to the network, 'please solve this standard computer vision task, but please also make the function of one of your inside simulated neural layers be as similar as possible to the function of the corresponding biological neural layer,'" DiCarlo explains. "We asked it to do both of those things as best it could." This forced the artificial neural circuits to find a different way to process visual information than the standard, computer vision approach, he says.

After training the artificial model with biological data, DiCarlo's team compared its activity to a similarly-sized neural network model trained without neural data, using the standard approach for computer vision. They found that the new, biologically informed model IT layer was—as instructed—a better match for IT neural data. That is, for every image tested, the population of artificial IT neurons in the model responded more similarly to the corresponding population of biological IT neurons.

The researchers also found that the model IT was also a better match to IT neural data collected from another monkey, even though the model had never seen data from that animal, and even when that comparison was evaluated on that monkey's IT responses to new images. This indicated that the team's new, "neurally aligned" computer model may be

an improved model of the neurobiological function of the primate IT cortex—an interesting finding, given that it was previously unknown whether the amount of neural data that can be currently collected from the primate visual system is capable of directly guiding model development.

With their new computer model in hand, the team asked whether the "IT neural alignment" procedure also leads to any changes in the overall behavioral performance of the model. Indeed, they found that the neurally-aligned model was more human-like in its behavior—it tended to succeed in correctly categorizing objects in images for which humans also succeed, and it tended to fail when humans also fail.

Adversarial attacks

The team also found that the neurally aligned model was more resistant to "adversarial attacks" that developers use to test computer vision and AI systems. In computer vision, adversarial attacks introduce small distortions into images that are meant to mislead an artificial neural network.

"Say that you have an image that the model identifies as a cat. Because you have the knowledge of the internal workings of the model, you can then design very small changes in the image so that the model suddenly thinks it's no longer a cat," DiCarlo explains.

These minor distortions don't typically fool humans, but computer vision models struggle with these alterations. A person who looks at the subtly distorted cat still reliably and robustly reports that it's a cat. But standard computer vision models are more likely to mistake the cat for a dog, or even a tree.

"There must be some internal differences in the way our brains process

images that lead to our vision being more resistant to those kinds of attacks," DiCarlo says. And indeed, the team found that when they made their model more neurally aligned, it became more robust, correctly identifying more images in the face of adversarial attacks. The model could still be fooled by stronger "attacks," but so can people, DiCarlo says. His team is now exploring the limits of adversarial robustness in humans.

A few years ago, DiCarlo's team found they could also improve a model's resistance to adversarial attacks by designing the first layer of the artificial network to emulate the early visual processing layer in the brain. One key next step is to combine such approaches—making new models that are simultaneously neurally aligned at multiple visual processing layers.

The new work is further evidence that an exchange of ideas between neuroscience and computer science can drive progress in both fields. "Everybody gets something out of the exciting virtuous cycle between natural/biological intelligence and artificial intelligence," DiCarlo says. "In this case, computer vision and AI researchers get new ways to achieve robustness, and neuroscientists and cognitive scientists get more accurate mechanistic models of human vision."

More information: Aligning Model and Macaque Inferior Temporal Cortex Representations Improves Model-to-Human Behavioral Alignment and Adversarial Robustness. openreview.net/attachment?id=SMYdcXjJh1q&name=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.