TechXplore

New tool explains how AI 'sees' images and why it might mistake an astronaut for a shovel

June 28 2023

Credit: Brown University

Why is it that artificial intelligence systems can outperform humans on some visual tasks, like facial recognition, but make egregious errors on others—such as classifying an image of an astronaut as a shovel?

Like the human brain, AI systems rely on strategies for processing and classifying images. And like the human brain, little is known about the precise nature of those processes. Scientists at Brown University's Carney Institute for Brain Science are making strides in understanding both systems, publishing a recent paper that helps to explain computer vision in a way the researchers say is accessible as well as more useful than previous models.

"Both the human brain and the deep neural networks that power AI systems are referred to as black boxes because we don't know exactly what goes on inside," said Thomas Serre, a Brown professor of cognitive, linguistic and psychological sciences and computer science. "The work that we do at Carney's Center for Computational Brain Science is trying to understand and characterize brain mechanisms related to learning, vision, and all kinds of things, and highlighting the similarities and differences with AI systems."

Deep neural networks use learning algorithms to process images, Serre said. They are trained on massive sets of data, such as ImageNet, which has over a million images culled from the web organized into thousands of object categories. The training mainly involves feeding data to the AI system, he explained.

"We don't tell AI systems how to process images—for example, what information to extract from the images to be able to classify them," Serre said. "The AI system discovers its own strategy. Then computer scientists evaluate the accuracy of what they do after they've been trained—for example, maybe the system achieves 90% accuracy on discriminating between a thousand image categories."

Serre collaborated with Brown Ph.D. candidate Thomas Fel and other computer scientists to develop a tool that allows users to pry open the lid of the black box of deep neural networks and illuminate what types of strategies AI systems use to process images. The project, called CRAFT—for Concept Recursive Activation FacTorization for Explainability—was a joint project with the Artificial and Natural Intelligence Toulouse Institute, where Fel is currently based. It was presented this month at the IEEE/CVF Conference on Computer Vision and Pattern Recognition in Vancouver, Canada.

Serre shared how CRAFT reveals how AI "sees" images and explained the crucial importance of understanding how the computer vision system differs from the human one.

What does CRAFT show about the way AI processes images?

CRAFT provides an interpretation of the complex and high-dimensional visual representations of objects learned by neural networks, leveraging modern machine learning tools to make them more understandable to humans. This leads to a representation of the key visual concepts used by neural networks to classify objects. As an example, let's think about a type of freshwater fish called a tench. We built a website that allows people to browse and visualize these concepts. Using the website, one can see that AI system's concept of a tench includes sets of fish fins, heads, tails, eyeballs and more.

These concepts also reveal that deep networks sometimes pick up on biases in datasets. One of the concepts associated with the tench, for example, is the face of a white male, because there are many photos online of sports fishermen holding fish that look like tench. (Yet the system can still distinguish a man from a fish.) In another example, the predominant concept associated with a soccer ball in neural networks is the presence of soccer players on the field. This is likely because the majority of internet images featuring soccer balls also include individual players rather than solely the ball itself.

How does the CRAFT method differ from other ways of understanding computer vision?

One way to explain AI vision is through what's called attribution methods, which employ heatmaps to identify the most influential regions of an image that impact AI decisions. However, these methods mainly focus on the most prominent regions of an image—revealing "where" the model looks, but failing to explain "what" the model sees in those areas.

The improvement over previous methods that Thomas Fel has introduced with CRAFT is to not only identify what concepts the system is using to piece together an image or what the model sees in those areas, but also how the system is ranking those concepts. In the tench example, the fish torso corresponds to 60% of the entire weight of the concept of a tench. So we can learn how much weight the AI system is placing on those subconcepts. In other words, it is more likely to classify an image with a tench torso as a fish than it is to classify an image with a white male as a fish.

How can CRAFT explain why AI vision systems sometimes make bizarre mistakes?

In our paper, we use CRAFT to explain a classic AI mistake: how an image of an astronaut was incorrectly classified as a shovel by an AI system trained on ImageNet. Here's what happened: A heatmap generated by a classic attribution method showed that the system was looking at the middle of the image in a shape of a shovel. The CRAFT approach highlighted the two most influential concepts that drove the decision along with their corresponding locations.

CRAFT suggested that the neural network arrived at its decision because it identified the concept of "dirt" commonly found in members of the image class "shovel" and the concept of "ski pants" typically worn by people clearing snow from their driveway with a shovel. It should have identified the correct concept of astronaut's pants, but that image of pants was probably never seen during the training process, so the system wasn't able to make that connection.

Why is it so important to understand the details of how a computer sees images?

First, it helps improve the accuracy and performance of vision-based tools like facial recognition. It makes AI systems more trustworthy because we can understand the visual strategy they're using. And it also helps make them safer against cyber-attacks. Take, for example, the concept of adversarial attacks. The fact is that one can make tiny alterations on images such as by changing pixel intensities in ways that are barely perceptible to humans yet that will be sufficient to completely fool the AI system.

In one crucially important example, researchers have shown that by simply adding a few stickers in a particular pattern to a stop sign, they can fool a self-driving vehicle to process it as a speed limit sign so that it will speed up instead of slow down and stop. That would create complete chaos. So we need to be able to understand why and how these types of attacks work on AI in order to be able to safeguard against them.

What can AI vision systems teach us about human vision systems?

We have learned that there's something fundamentally different from the way these neural networks process images compared to the human brain—the human brain would not process a stop sign with some black stickers on it as a speed limit sign. Yet these are incredibly well-engineered systems that sometimes even outperform humans, like in facial recognition tasks. A lot of the work we do in our lab is to compare what's similar and what's different about these systems.

Whenever we're able to find limitations of AI systems, we then go to neuroscience and ask, "What is the brain mechanism that is missing in AI systems that we know is playing a key role in humans' ability to solve this task robustly and efficiently?" And then we build machine learning abstractions of this mechanism and inject them in an AI system of the neural network. We find that once we endow AI systems with human brain-inspired mechanisms, they perform much better: They are more robust, more efficient at learning, and are more accurate with less training.

On the neuroscience side, this research helps us better understand the human brain and how these differences between humans and AI systems help humans, and we can also validate our ideas more easily and more safely than we could in a human brain. It's very hard to understand how the brain processes visual information. There have been methods developed to understand how neurons work and what they do, and with AI systems, we can now test those theories and see if we're right.

The synergies go both ways: neuroscience gives us good inspiration for improving AI. But the fact that we do improve AI from those neuroscience mechanisms is also a way to validate the discoveries made in neuroscience and to identify key mechanisms of general intelligence, visual intelligence and more.

We hear a lot about the worries about AI systems that are too human-like. But it sounds like when it comes

to vision, that's a good thing, isn't it?

In many cases we found very significant benefits for humans in aligning the AI vision system and the human vision system: The models that are made to be more human-like become more trustworthy, reliable, resilient to attacks and less likely to do what you don't want them to do.

What are the next steps in this research?

It's interesting to see how AI systems categorize natural objects, but I think what's next will be to use what we've learned about AI and human vision to help AI systems tackle big problems in science that humans are unable to solve right now—like in cancer diagnostics, for example, or in fossil recognition, or in space exploration. That will be really exciting.

More information: Thomas Fel et al, CRAFT: Concept Recursive Activation FacTorization for Explainability (2023)

Provided by Brown University

Citation: New tool explains how AI 'sees' images and why it might mistake an astronaut for a shovel (2023, June 28) retrieved 9 May 2024 from https://techxplore.com/news/2023-06-tool-ai-images-astronaut-shovel.html