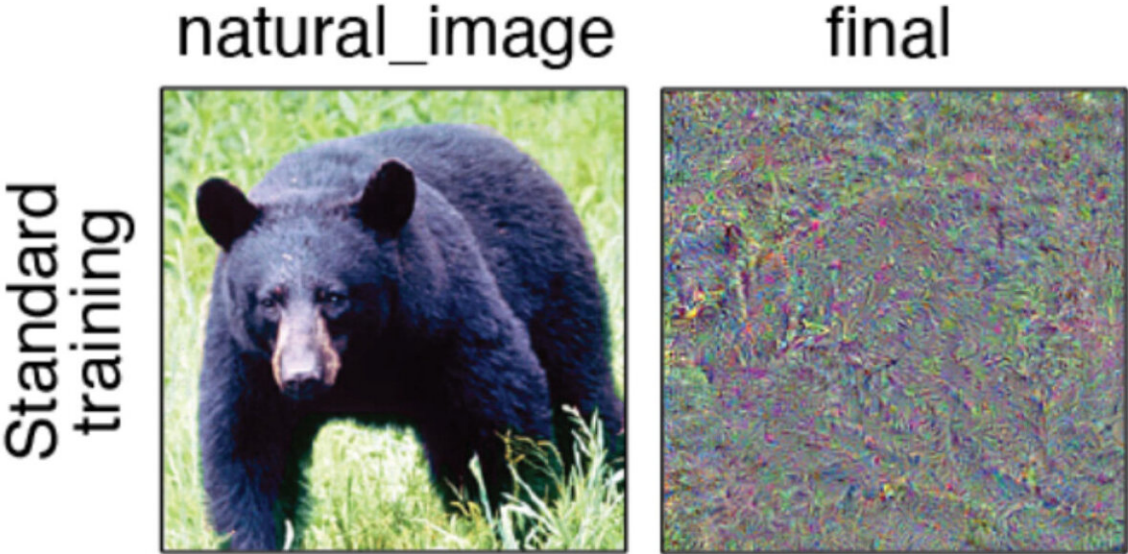# Study: Deep neural networks don't see the world the way we do

October 16 2023, by Anne Trafton

When these neural networks were asked to generate an image or a word that they would put in the same category as a specific input, such as a picture of a bear, most of what they produced was unrecognizable to human observers. On the right is an example of what the model categorized as "bear.". Credit: MIT researchers

Human sensory systems are very good at recognizing objects that we see or words that we hear, even if the object is upside down or the word is spoken by a voice we've never heard.

Computational models known as [deep neural networks](link) can be trained to do the same thing, correctly identifying an image of a dog regardless of what color its fur is, or a word regardless of the pitch of the speaker's voice. However, a new study from MIT neuroscientists has found that these models often also respond the same way to images or words that have no resemblance to the target.

When these neural networks were used to generate an image or a word that they responded to in the same way as a specific natural input, such as a picture of a bear, most of them generated images or sounds that were unrecognizable to human observers. This suggests that these models build up their own idiosyncratic "invariances"—meaning that they respond the same way to stimuli with very different features.

The findings offer a new way for researchers to evaluate how well these models mimic the organization of human sensory perception, says Josh McDermott, an associate professor of brain and cognitive sciences at MIT and a member of MIT's McGovern Institute for Brain Research and Center for Brains, Minds, and Machines.

"This paper shows that you can use these models to derive unnatural signals that end up being very diagnostic of the representations in the model," says McDermott, who is the senior author of the study. "This test should become part of a battery of tests that we as a field are using to evaluate models."

Jenelle Feather Ph.D. '22, who is now a research fellow at the Flatiron Institute Center for Computational Neuroscience, is the lead author of the open-access paper, which [appears in *Nature Neuroscience*](link). Guillaume

Leclerc, an MIT graduate student, and Aleksander Mądry, the Cadence Design Systems Professor of Computing at MIT, are also authors of the paper.

Different perceptions

In recent years, researchers have trained deep neural networks that can analyze millions of inputs (sounds or images) and learn common features that allow them to classify a target word or object roughly as accurately as humans do. These models are currently regarded as the leading models of biological sensory systems.

It is believed that when the human sensory system performs this kind of classification, it learns to disregard features that aren't relevant to an object's core identity, such as how much light is shining on it or what angle it's being viewed from. This is known as invariance, meaning that objects are perceived to be the same even if they show differences in those less important features.

"Classically, the way that we have thought about sensory systems is that they build up invariances to all those sources of variation that different examples of the same thing can have," Feather says. "An organism has to recognize that they're the same thing even though they show up as very different sensory signals."

The researchers wondered if deep neural networks that are trained to perform classification tasks might develop similar invariances. To try to answer that question, they used these models to generate stimuli that produce the same kind of response within the model as an example stimulus given to the model by the researchers.

They term these stimuli "model metamers," reviving an idea from classical perception research whereby stimuli that are indistinguishable

to a system can be used to diagnose its invariances. The concept of metamers was originally developed in the study of human perception to describe colors that look identical even though they are made up of different wavelengths of light.

To their surprise, the researchers found that most of the images and sounds produced in this way looked and sounded nothing like the examples that the models were originally given. Most of the images were a jumble of random-looking pixels, and the sounds resembled unintelligible noise. When researchers showed the images to human observers, in most cases the humans did not classify the images synthesized by the models in the same category as the original target example.

"They're really not recognizable at all by humans. They don't look or sound natural and they don't have interpretable features that a person could use to classify an object or word," Feather says.

The findings suggest that the models have somehow developed their own invariances that are different from those found in human perceptual systems. This causes the models to perceive pairs of stimuli as being the same despite their being wildly different to a human.

Idiosyncratic invariances

The researchers found the same effect across many different vision and auditory models. However, each of these models appeared to develop their own unique invariances. When metamers from one model were shown to another model, the metamers were just as unrecognizable to the second model as they were to human observers.

"The key inference from that is that these models seem to have what we call idiosyncratic invariances," McDermott says. "They have learned to

be invariant to these particular dimensions in the stimulus space, and it's model-specific, so other models don't have those same invariances."

The researchers also found that they could induce a model's metamers to be more recognizable to humans by using an approach called adversarial training. This approach was originally developed to combat another limitation of object recognition models, which is that introducing tiny, almost imperceptible changes to an image can cause the model to misrecognize it.

The researchers found that adversarial training, which involves including some of these slightly altered images in the training data, yielded models whose metamers were more recognizable to humans, though they were still not as recognizable as the original stimuli. This improvement appears to be independent of the training's effect on the models' ability to resist adversarial attacks, the researchers say.

"This particular form of training has a big effect, but we don't really know why it has that effect," Feather says. "That's an area for future research."

Analyzing the metamers produced by computational models could be a useful tool to help evaluate how closely a computational model mimics the underlying organization of human sensory perception systems, the researchers say.

"This is a behavioral test that you can run on a given model to see whether the invariances are shared between the model and human observers," Feather says. "It could also be used to evaluate how idiosyncratic the invariances are within a given model, which could help uncover potential ways to improve our models in the future."

 More information: Model metamers reveal divergent invariances

between biological and artificial neural networks, *Nature Neuroscience* (2023). DOI: 10.1038/s41593-023-01442-0. www.nature.com/articles/s41593-023-01442-0

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