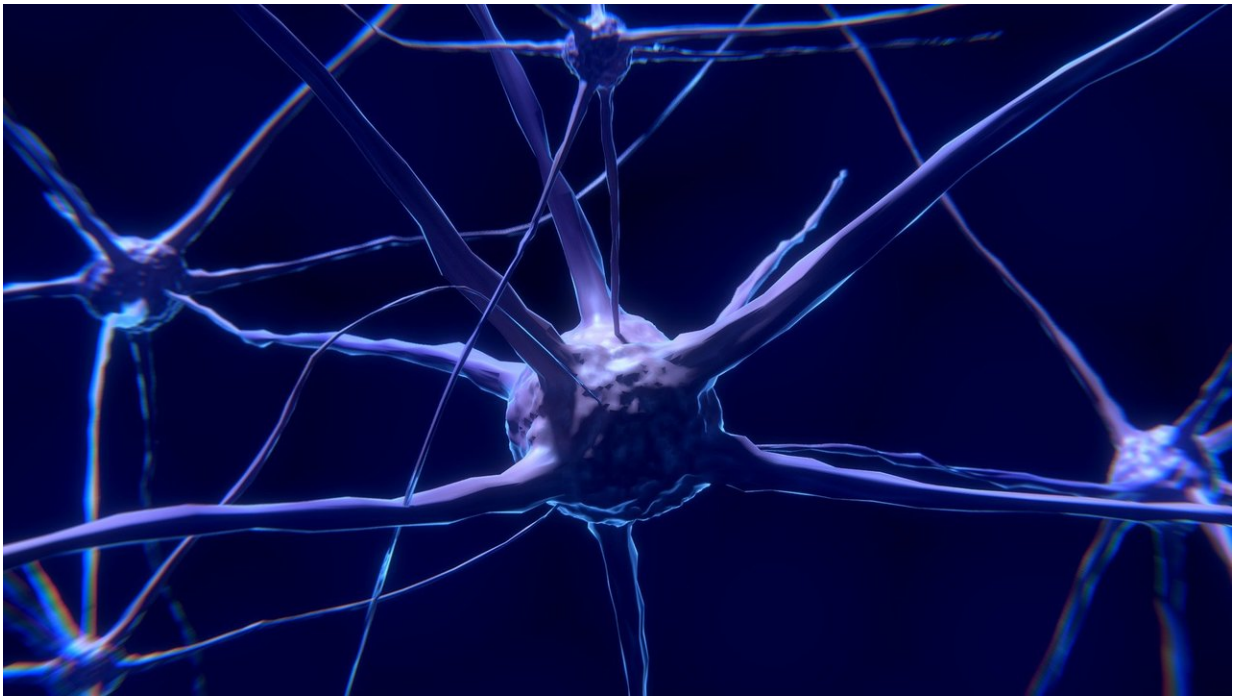
Differences between deep neural networks and human perception

December 13 2019, by Kenneth I. Blum



Credit: CC0 Public Domain

When your mother calls your name, you know it's her voice—no matter the volume, even over a poor cell phone connection. And when you see her face, you know it's hers—if she is far away, if the lighting is poor, or if you are on a bad FaceTime call. This robustness to variation is a hallmark of human perception. On the other hand, we are susceptible to illusions: We might fail to distinguish between sounds or images that are,

in fact, different. Scientists have explained many of these illusions, but we lack a full understanding of the invariances in our auditory and visual systems.

Deep neural networks also have performed speech recognition and image classification tasks with impressive robustness to variations in the auditory or visual stimuli. But are the invariances learned by these models similar to the invariances learned by human perceptual systems? A group of MIT researchers has discovered that they are different. They presented their findings yesterday at the 2019 Conference on Neural Information Processing Systems.

The researchers made a novel generalization of a classical concept: "metamers"—physically distinct stimuli that generate the same perceptual effect. The most famous examples of metamer stimuli arise because most people have three different types of cones in their retinae, which are responsible for color vision. The perceived color of any single wavelength of light can be matched exactly by a particular combination of three lights of different colors—for example, red, green, and blue lights. Nineteenth-century scientists inferred from this observation that humans have three different types of bright-light detectors in our eyes. This is the basis for electronic color displays on all of the screens we stare at every day. Another example in the visual system is that when we fix our gaze on an object, we may perceive surrounding visual scenes that differ at the periphery as identical. In the auditory domain, something analogous can be observed. For example, the "textural" sound of two swarms of insects might be indistinguishable, despite differing in the acoustic details that compose them, because they have similar aggregate statistical properties. In each case, the metamers provide insight into the mechanisms of perception, and constrain models of the human visual or auditory systems.

In the current work, the researchers randomly chose natural images and

sound clips of spoken words from standard databases, and then synthesized sounds and images so that deep neural networks would sort them into the same classes as their natural counterparts. That is, they generated physically distinct stimuli that are classified identically by models, rather than by humans. This is a new way to think about metamers, generalizing the concept to swap the role of computer models for human perceivers. They therefore called these synthesized stimuli "model metamers" of the paired natural stimuli. The researchers then tested whether humans could identify the words and images.

"Participants heard a short segment of speech and had to identify from a list of words which word was in the middle of the clip. For the natural audio this task is easy, but for many of the model metamers humans had a hard time recognizing the sound," explains first-author Jenelle Feather, a graduate student in the MIT Department of Brain and Cognitive Sciences (BCS) and a member of the Center for Brains, Minds, and Machines (CBMM). That is, humans would not put the synthetic stimuli in the same class as the spoken word "bird" or the image of a bird. In fact, model metamers generated to match the responses of the deepest layers of the model were generally unrecognizable as words or images by human subjects.

Josh McDermott, associate professor in BCS and investigator in CBMM, makes the following case: "The basic logic is that if we have a good model of human perception, say of speech recognition, then if we pick two sounds that the model says are the same and present these two sounds to a human listener, that human should also say that the two sounds are the same. If the human listener instead perceives the stimuli to be different, this is a clear indication that the representations in our model do not match those of human perception."

Joining Feather and McDermott on the paper are Alex Durango, a post-baccalaureate student, and Ray Gonzalez, a research assistant, both in

BCS.

There is another type of failure of deep networks that has received a lot of attention in the media: adversarial examples (see, for example, "Why did my classifier just mistake a turtle for a rifle?"). These are stimuli that appear similar to humans but are misclassified by a model network (by design—they are constructed to be misclassified). They are complementary to the stimuli generated by Feather's group, which sound or appear different to humans but are designed to be co-classified by the model network. The vulnerabilities of model networks exposed to adversarial attacks are well-known—face-recognition software might mistake identities; automated vehicles might not recognize pedestrians.

The importance of this work lies in improving models of perception beyond deep networks. Although the standard adversarial examples indicate differences between deep networks and human perceptual systems, the new stimuli generated by the McDermott group arguably represent a more fundamental model failure—they show that generic examples of stimuli classified as the same by a deep network produce wildly different percepts for humans.

The team also figured out ways to modify the model networks to yield metamers that were more plausible sounds and images to humans. As McDermott says, "This gives us hope that we may be able to eventually develop models that pass the metamer test and better capture human invariances."

"Model metamers demonstrate a significant failure of present-day neural networks to match the invariances in the human visual and auditory systems," says Feather, "We hope that this work will provide a useful behavioral measuring stick to improve model representations and create better models of human sensory systems."