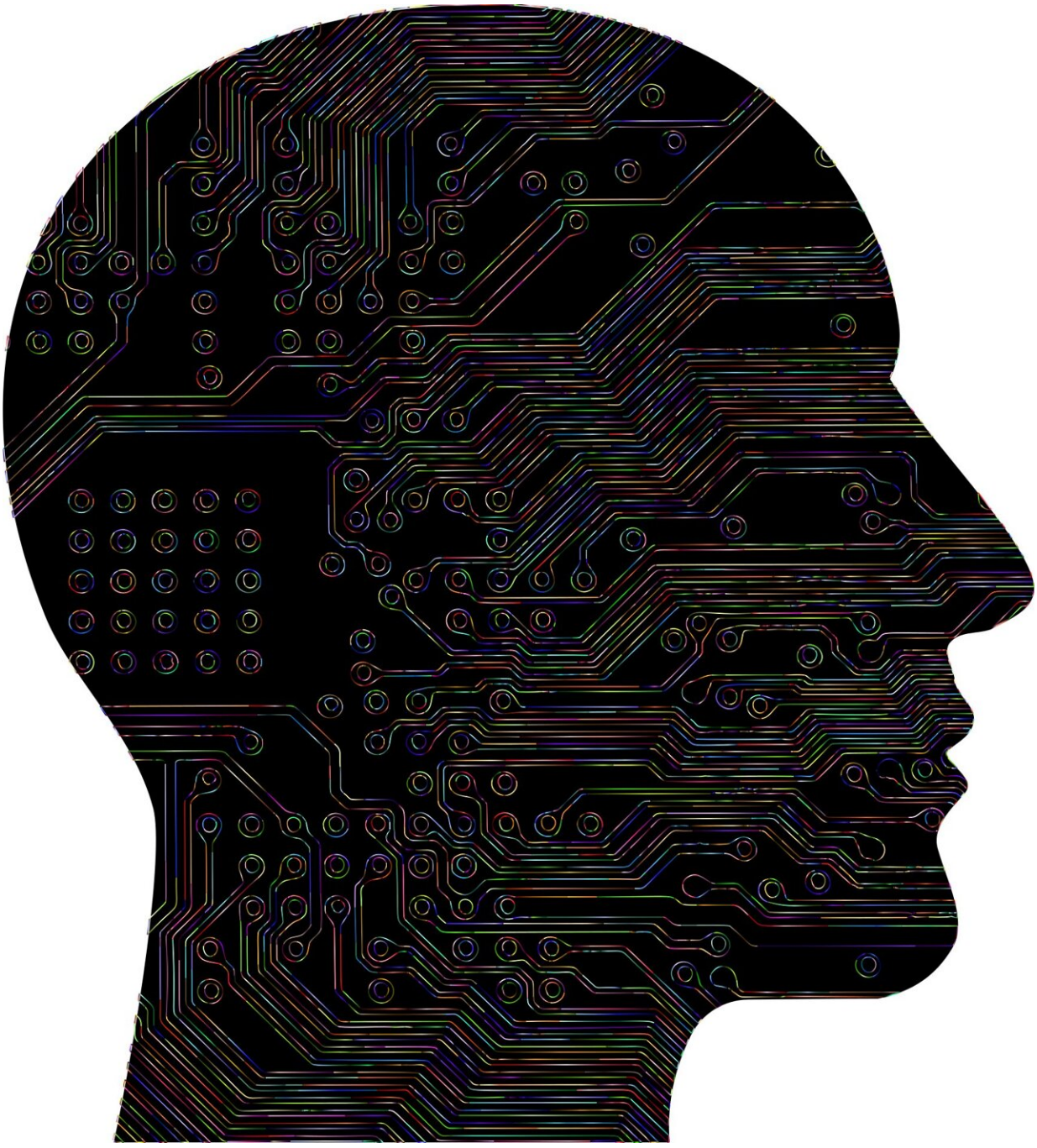
The brain may learn about the world the same way some computational models do

October 30 2023, by Anne Trafton



Credit: Pixabay/CC0 Public Domain

To make our way through the world, our brain must develop an intuitive

understanding of the physical world around us, which we then use to interpret sensory information coming into the brain.

How does the brain develop that intuitive understanding? Many scientists believe that it may use a process similar to what's known as "self-supervised learning." This type of machine learning, originally developed as a way to create more efficient models for computer vision, allows computational models to learn about visual scenes based solely on the similarities and differences between them, with no labels or other information.

A pair of studies from researchers at the K. Lisa Yang Integrative Computational Neuroscience (ICoN) Center at MIT offers new evidence supporting this hypothesis. The researchers found that when they trained models known as neural networks using a particular type of self-supervised learning, the resulting models generated activity patterns very similar to those seen in the brains of animals that were performing the same tasks as the models.

The findings suggest that these models are able to learn representations of the physical world that they can use to make accurate predictions about what will happen in that world, and that the mammalian brain may be using the same strategy, the researchers say.

"The theme of our work is that AI designed to help build better robots ends up also being a framework to better understand the brain more generally," says Aran Nayebi, a postdoc in the ICoN Center. "We can't say if it's the whole brain yet, but across scales and disparate brain areas, our results seem to be suggestive of an organizing principle."

Nayebi is the lead author of one of the studies, co-authored with Rishi Rajalingham, a former MIT postdoc now at Meta Reality Labs, and senior authors Mehrdad Jazayeri, an associate professor of brain and

cognitive sciences and a member of the McGovern Institute for Brain Research; and Robert Yang, an assistant professor of brain and cognitive sciences and an associate member of the McGovern Institute.

Ila Fiete, director of the ICoN Center, a professor of brain and cognitive sciences, and an associate member of the McGovern Institute, is the senior author of the other study, which was co-led by Mikail Khona, an MIT graduate student, and Rylan Schaeffer, a former senior research associate at MIT.

Both studies will be presented at the 2023 Conference on Neural Information Processing Systems (NeurIPS) in December.

Modeling the physical world

Early models of computer vision mainly relied on supervised learning. Using this approach, models are trained to classify images that are each labeled with a name—cat, car, etc. The resulting models work well, but this type of training requires a great deal of human-labeled data.

To create a more efficient alternative, in recent years researchers have turned to models built through a technique known as contrastive self-supervised learning. This type of learning allows an algorithm to learn to classify objects based on how similar they are to each other, with no external labels provided.

"This is a very powerful method because you can now leverage very large modern data sets, especially videos, and really unlock their potential," Nayebi says. "A lot of the modern AI that you see now, especially in the last couple years with ChatGPT and GPT-4, is a result of training a self-supervised objective function on a large-scale dataset to obtain a very flexible representation."

These types of models, also called neural networks, consist of thousands or millions of processing units connected to each other. Each node has connections of varying strengths to other nodes in the network. As the network analyzes huge amounts of data, the strengths of those connections change as the network learns to perform the desired task.

As the model performs a particular task, the activity patterns of different units within the network can be measured. Each unit's activity can be represented as a firing pattern, similar to the firing patterns of neurons in the brain. Previous work from Nayebi and others has shown that self-supervised models of vision generate activity similar to that seen in the visual processing system of mammalian brains.

In both of the new NeurIPS studies, the researchers set out to explore whether self-supervised computational models of other cognitive functions might also show similarities to the mammalian brain. In the study led by Nayebi, the researchers trained self-supervised models to predict the future state of their environment across hundreds of thousands of naturalistic videos depicting everyday scenarios.

"For the last decade or so, the dominant method to build neural network models in cognitive neuroscience is to train these networks on individual cognitive tasks. But models trained this way rarely generalize to other tasks," Yang says. "Here we test whether we can build models for some aspect of cognition by first training on naturalistic data using self-supervised learning, then evaluating in lab settings."

Once the model was trained, the researchers had it generalize to a task they call "Mental-Pong." This is similar to the video game Pong, where a player moves a paddle to hit a ball traveling across the screen. In the Mental-Pong version, the ball disappears shortly before hitting the paddle, so the player has to estimate its trajectory in order to hit the ball.

The researchers found that the model was able to track the hidden ball's trajectory with accuracy similar to that of neurons in the mammalian brain, which had been shown in a previous study by Rajalingham and Jazayeri to simulate its trajectory—a cognitive phenomenon known as "mental simulation." Furthermore, the neural activation patterns seen within the model were similar to those seen in the brains of animals as they played the game—specifically, in a part of the brain called the dorsomedial frontal cortex. No other class of computational model has been able to match the biological data as closely as this one, the researchers say.

"There are many efforts in the machine learning community to create artificial intelligence," Jazayeri says. "The relevance of these models to neurobiology hinges on their ability to additionally capture the inner workings of the brain. The fact that Aran's model predicts neural data is really important as it suggests that we may be getting closer to building artificial systems that emulate natural intelligence."

Navigating the world

The study led by Khona, Schaeffer, and Fiete focused on a type of specialized neurons known as grid cells. These cells, located in the entorhinal cortex, help animals to navigate, working together with place cells located in the hippocampus.

While place cells fire whenever an animal is in a specific location, grid cells fire only when the animal is at one of the vertices of a triangular lattice. Groups of grid cells create overlapping lattices of different sizes, which allows them to encode a large number of positions using a relatively small number of cells.

In recent studies, researchers have trained supervised neural networks to mimic grid cell function by predicting an animal's next location based on

its starting point and velocity, a task known as path integration. However, these models hinged on access to privileged information about absolute space at all times—information that the animal does not have.

Inspired by the striking coding properties of the multiperiodic grid-cell code for space, the MIT team trained a contrastive self-supervised model to both perform this same path integration task and represent space efficiently while doing so. For the training data, they used sequences of velocity inputs. The model learned to distinguish positions based on whether they were similar or different—nearby positions generated similar codes, but further positions generated more different codes.

"It's similar to training models on images, where if two images are both heads of cats, their codes should be similar, but if one is the head of a cat and one is a truck, then you want their codes to repel," Khona says. "We're taking that same idea but applying it to spatial trajectories."

Once the model was trained, the researchers found that the activation patterns of the nodes within the model formed several lattice patterns with different periods, very similar to those formed by grid cells in the brain.

"What excites me about this work is that it makes connections between mathematical work on the striking information-theoretic properties of the grid cell code and the computation of path integration," Fiete says. "While the mathematical work was analytic—what properties does the grid cell code possess?—the approach of optimizing coding efficiency through self-supervised learning and obtaining grid-like tuning is synthetic: It shows what properties might be necessary and sufficient to explain why the brain has grid cells."

More information: Aran Nayebi et al, Neural Foundations of Mental Simulation: Future Prediction of Latent Representations on Dynamic

Scenes, *arXiv* (2023). DOI: 10.48550/arxiv.2305.11772

Self-Supervised Learning of Representations for Space Generates Multi-Modular Grid Cells, neurips.cc/virtual/2023/poster/72628