Artificial intelligence sheds light on how the brain processes language – RESEARCHNEWS October 28, 2021

The most recent generation of predictive language models also appears to learn something about the underlying meaning of language.

CAMBRIDGE, MASS.- In the past few years, artificial intelligence models of language have become very good at certain tasks. Most notably, they excel at predicting the next word in a string of text; this technology helps search engines and texting apps predict the next word you are going to type.

The most recent generation of predictive language models also appears to learn something about the underlying meaning of language. These models can not only predict the word that comes next, but also perform tasks that seem to require some degree of genuine understanding, such as question answering, document summarization, and story completion.

Such models were designed to optimize performance for the specific function of predicting text, without attempting to mimic anything about how the human brain performs this task or understands language. But a new study from MIT neuroscientists suggests the underlying function of these models resembles the function of language-processing centers in the human brain.

Computer models that perform well on other types of language tasks do not show this similarity to the human brain, offering evidence that the human brain may use next-word prediction to drive language processing.

"The better the model is at predicting the next word, the more closely it fits the human brain," says Nancy Kanwisher, the Walter A. Rosenblith Professor of Cognitive Neuroscience, a member of MIT's McGovern Institute for Brain Research and Center for Brains, Minds, and Machines (CBMM), and an author of the new study. "It's amazing that the models fit so well, and it very indirectly suggests that maybe what the human language system is doing is predicting what's going to happen next."

Joshua Tenenbaum, a professor of computational cognitive science at MIT and a member of CBMM and MIT's Artificial Intelligence Laboratory (CSAIL); and Evelina Fedorenko, the Frederick A. and Carole J. Middleton Career Development Associate Professor of Neuroscience and a member of the McGovern Institute, are the senior authors of the study, which appears this week in the Proceedings of the National Academy of Sciences. Martin Schrimpf, an MIT graduate student who works in CBMM, is the first author of the paper.

Making predictions

The new, high-performing next-word prediction models belong to a class of models called deep neural networks. These networks contain computational "nodes" that form connections of varying strength, and layers that pass information between each other in prescribed ways.

Over the past decade, scientists have used deep neural networks to create models of vision that can recognize objects as well as the primate brain does. Research at MIT has also shown that the underlying function of visual object recognition models matches the organization of the primate visual cortex, even though those computer models were not specifically designed to mimic the brain.

In the new study, the MIT team used a similar approach to compare language-processing centers in the human brain with language-processing models. The researchers analyzed 43 different language models, including several that are optimized for next-word prediction. These include a model called GPT-3 (Generative Pre-trained Transformer 3), which, given a prompt, can generate text similar to what a human would produce. Other models were designed to perform different language tasks, such as filling in a blank in a sentence.

As each model was presented with a string of words, the researchers measured the activity of the nodes that make up the network. They then compared these patterns to activity in the human brain, measured in subjects performing three language tasks: listening to stories, reading sentences one at a time, and reading sentences in which one word is revealed at a time. These human datasets included functional magnetic resonance (fMRI) data and intracranial electrocorticographic measurements taken in people undergoing brain surgery for epilepsy.

They found that the best-performing next-word prediction models had activity patterns that very closely resembled those seen in the human brain. Activity in those same models was also highly correlated with measures of human behavioral measures such as how fast people were able to read the text.

“We found that the models that predict the neural responses well also tend to best predict human behavior responses, in the form of reading times. And then both of these are explained by the model performance on next-word prediction. This triangle really connects everything together,” Schrimpf says.

“A key takeaway from this work is that language processing is a highly constrained problem: The best solutions to it that AI engineers have created end up being similar, as this paper shows, to the solutions found by the evolutionary process that created the human brain. Since the AI network didn't seek to mimic the brain directly — but does end up looking brain-like — this suggests that, in a sense, a kind of convergent evolution has occurred between AI and nature,” says Daniel Yamins, an assistant professor of psychology and computer science at Stanford University, who was not involved in the study.

Game changer

One of the key computational features of predictive models such as GPT-3 is an element known as a forward one-way predictive transformer. This kind of transformer is able to make predictions of what is going to come next, based on previous sequences. A significant feature of this transformer is that it can make predictions based on a very long prior context (hundreds of words), not just the last few words.

Scientists have not found any brain circuits or learning mechanisms that correspond to this type of processing, Tenenbaum says. However, the new findings are consistent with hypotheses that have been previously proposed that prediction is one of the key functions in language processing, he says.

“One of the challenges of language processing is the real-time aspect of it,” he says. “Language comes in, and you have to keep up with it and be able to make sense of it in real time.”

The researchers now plan to build variants of these language processing models to see how small changes in their architecture affect their performance and their ability to fit human neural data.

“For me, this result has been a game changer,” Fedorenko says. “It’s totally transforming my research program, because I would not have predicted that in my lifetime we would get to these computationally explicit models that capture enough about the brain so that we can actually leverage them in understanding how the brain works.”

The researchers also plan to try to combine these high-performing language models with some computer models Tenenbaum’s lab has previously developed that can perform other kinds of tasks such as constructing perceptual representations of the physical world.

“If we’re able to understand what these language models do and how they can connect to models which do things that are more like perceiving and thinking, then that can give us more integrative models of how things work in the brain,” Tenenbaum says. “This could take us toward better artificial intelligence models, as well as giving us better models of how more of the brain works and how general intelligence emerges, than we’ve had in the past.”

The research was funded by a Takeda Fellowship; the MIT Shoemaker Fellowship; the Semiconductor Research Corporation; the MIT Media Lab Consortia; the MIT Singleton Fellowship; the MIT Presidential Graduate Fellowship; the Friends of the McGovern Institute Fellowship; the MIT Center for Brains, Minds, and Machines, through the National Science Foundation; the National Institutes of Health; MIT’s Department of Brain and Cognitive Sciences; and the McGovern Institute.

Other authors of the paper are Idan Blank PhD ’16 and graduate students Greta Tuckute, Carina Kauf, and Eghbal Hosseini.

Reprinted with permission of MIT News