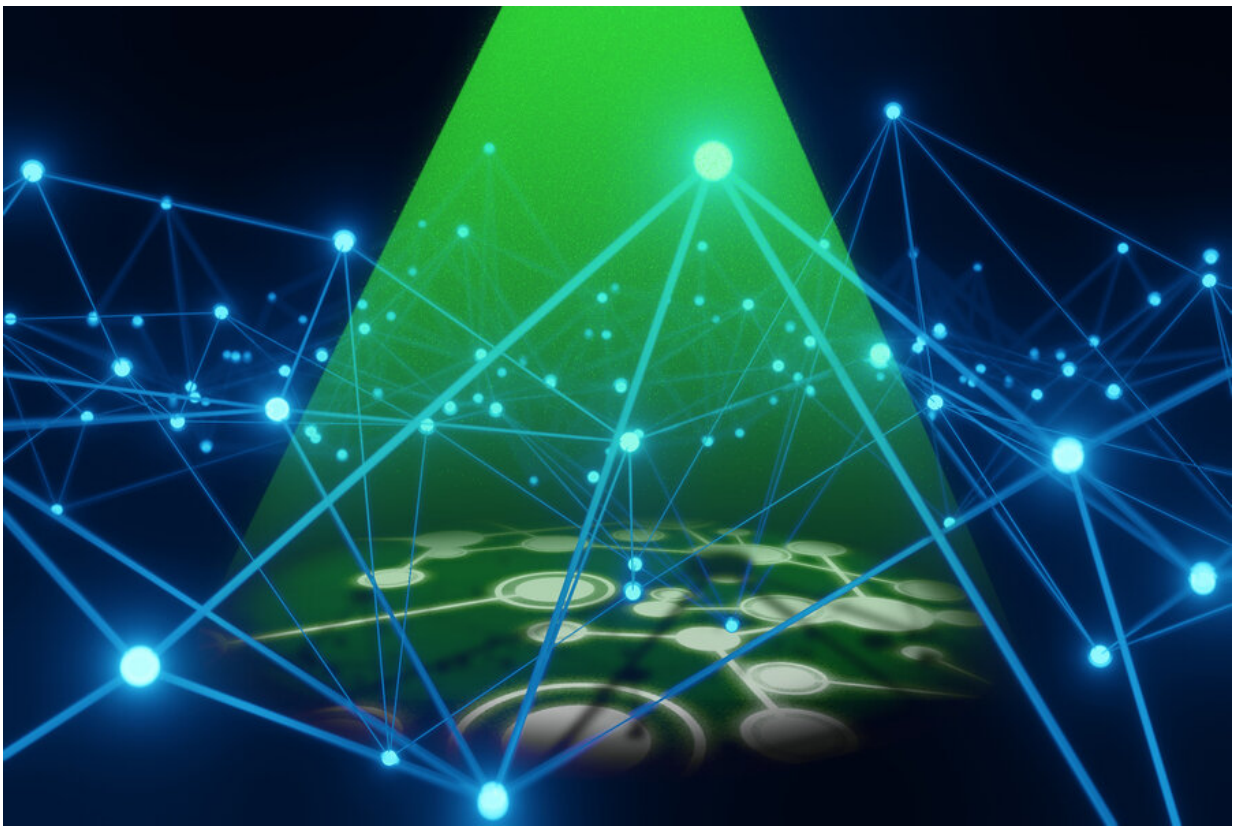
Study shows how large language models like GPT-3 can learn a new task from just a few examples

February 7 2023, by Adam Zewe



MIT researchers found that massive neural network models that are similar to large language models are capable of containing smaller linear models inside their hidden layers, which the large models could train to complete a new task using simple learning algorithms. Credit: Jose-Luis Olivares, MIT

Large language models like OpenAI's GPT-3 are massive neural networks that can generate human-like text, from poetry to programming code. Trained using troves of internet data, these machine-learning models take a small bit of input text and then predict the text that is likely to come next.

But that's not all these models can do. Researchers are exploring a curious phenomenon known as in-context learning, in which a large language model learns to accomplish a task after seeing only a few examples—despite the fact that it wasn't trained for that task. For instance, someone could feed the model several example sentences and their sentiments (positive or negative), then prompt it with a new sentence, and the model can give the correct sentiment.

Typically, a machine-learning model like GPT-3 would need to be retrained with new data for this new task. During this training process, the model updates its parameters as it processes new information to learn the task. But with in-context learning, the model's parameters aren't updated, so it seems like the model learns a new task without learning anything at all.

Scientists from MIT, Google Research, and Stanford University are striving to unravel this mystery. They studied models that are very similar to large language models to see how they can learn without updating parameters.

The researchers' theoretical results show that these massive neural network models are capable of containing smaller, simpler linear models buried inside them. The large model could then implement a simple learning algorithm to train this smaller, linear model to complete a new task, using only information already contained within the larger model. Its parameters remain fixed.

An important step toward understanding the mechanisms behind in-context learning, this research opens the door to more exploration around the learning algorithms these large models can implement, says Ekin Akyürek, a computer science graduate student and lead author of a paper exploring this phenomenon. With a better understanding of in-context learning, researchers could enable models to complete new tasks without the need for costly retraining.

"Usually, if you want to fine-tune these models, you need to collect domain-specific data and do some complex engineering. But now we can just feed it an input, five examples, and it accomplishes what we want. So in-context learning is a pretty exciting phenomenon," Akyürek says.

The paper is published on the *arXiv* preprint server.

Joining Akyürek on the paper are Dale Schuurmans, a research scientist at Google Brain and professor of computing science at the University of Alberta; as well as senior authors Jacob Andreas, the X Consortium Assistant Professor in the MIT Department of Electrical Engineering and Computer Science and a member of the MIT Computer Science and Artificial Intelligence Laboratory (CSAIL); Tengyu Ma, an assistant professor of computer science and statistics at Stanford; and Danny Zhou, principal scientist and research director at Google Brain. The research will be presented at the International Conference on Learning Representations.

A model within a model

In the machine-learning research community, many scientists have come to believe that large language models can perform in-context learning because of how they are trained, Akyürek says.

For instance, GPT-3 has hundreds of billions of parameters and was

trained by reading huge swaths of text on the internet, from Wikipedia articles to Reddit posts. So, when someone shows the model examples of a new task, it has likely already seen something very similar because its training dataset included text from billions of websites. It repeats patterns it has seen during training, rather than learning to perform new tasks.

Akyürek hypothesized that in-context learners aren't just matching previously seen patterns, but instead are actually learning to perform new tasks. He and others had experimented by giving these models prompts using synthetic data, which they could not have seen anywhere before, and found that the models could still learn from just a few examples. Akyürek and his colleagues thought that perhaps these neural network models have smaller machine-learning models inside them that the models can train to complete a new task.

"That could explain almost all of the learning phenomena that we have seen with these large models," he says.

To test this hypothesis, the researchers used a neural network model called a transformer, which has the same architecture as GPT-3, but had been specifically trained for in-context learning.

By exploring this transformer's architecture, they theoretically proved that it can write a linear model within its hidden states. A neural network is composed of many layers of interconnected nodes that process data. The hidden states are the layers between the input and output layers.

Their mathematical evaluations show that this linear model is written somewhere in the earliest layers of the transformer. The transformer can then update the linear model by implementing simple learning algorithms.

In essence, the model simulates and trains a smaller version of itself.

Probing hidden layers

The researchers explored this hypothesis using probing experiments, where they looked in the transformer's hidden layers to try and recover a certain quantity.

"In this case, we tried to recover the actual solution to the linear model, and we could show that the parameter is written in the hidden states. This means the linear model is in there somewhere," he says.

Building off this theoretical work, the researchers may be able to enable a transformer to perform in-context learning by adding just two layers to the neural network. There are still many technical details to work out before that would be possible, Akyürek cautions, but it could help engineers create models that can complete new tasks without the need for retraining with new data.

"The paper sheds light on one of the most remarkable properties of modern large language models—their ability to learn from data given in their inputs, without explicit training. Using the simplified case of linear regression, the authors show theoretically how models can implement standard learning algorithms while reading their input, and empirically which learning algorithms best match their observed behavior," says Mike Lewis, a research scientist at Facebook AI Research who was not involved with this work. "These results are a stepping stone to understanding how models can learn more complex tasks, and will help researchers design better training methods for language models to further improve their performance."

Moving forward, Akyürek plans to continue exploring in-context learning with functions that are more complex than the linear models

they studied in this work. They could also apply these experiments to large language models to see whether their behaviors are also described by simple learning algorithms. In addition, he wants to dig deeper into the types of pretraining data that can enable in-context learning.

"With this work, people can now visualize how these models can learn from exemplars. So, my hope is that it changes some people's views about in-context learning," Akyürek says. "These models are not as dumb as people think. They don't just memorize these tasks. They can learn new tasks, and we have shown how that can be done."

More information: Ekin Akyürek et al, What learning algorithm is in-context learning? Investigations with linear models, *arXiv* (2022). DOI: 10.48550/arxiv.2211.15661

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