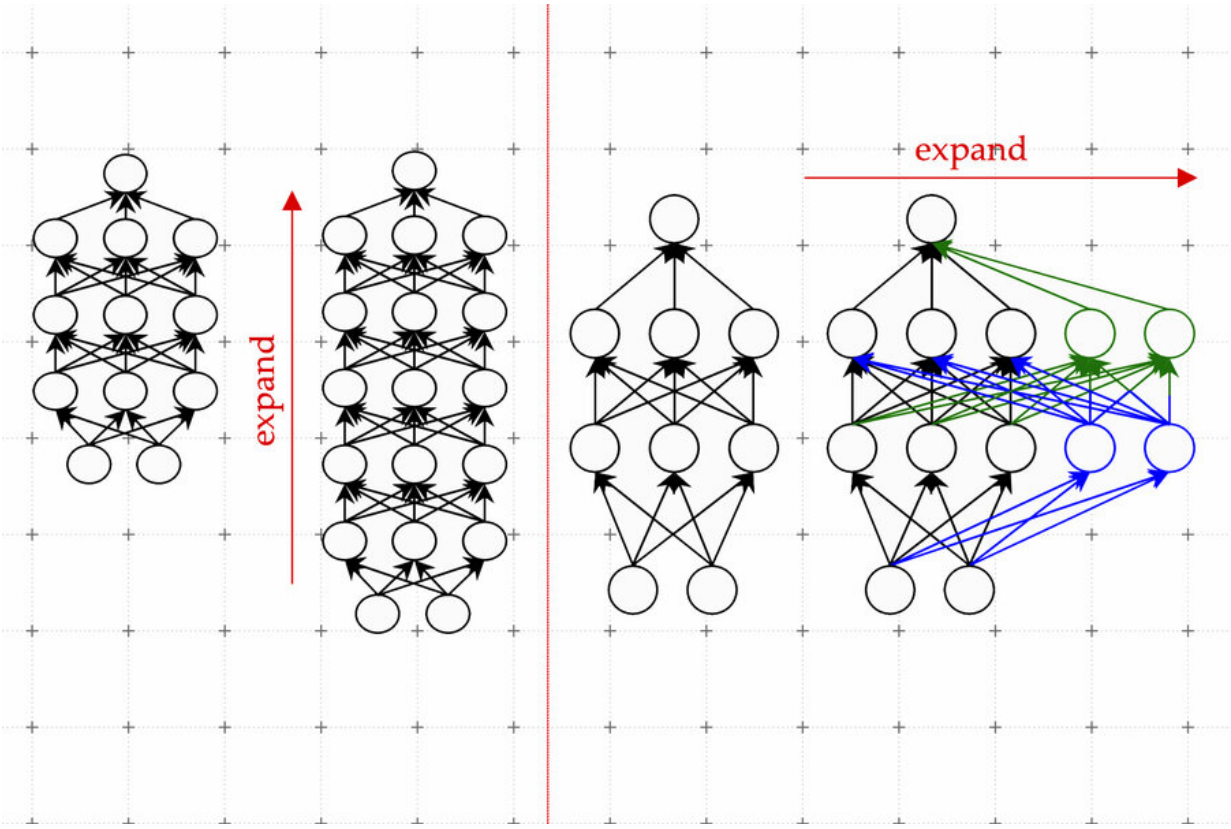
New LiGO technique accelerates training of large machine-learning models

March 22 2023, by Adam Zewe



The framework developed by the researchers accelerates training of a new, larger neural network model by using the weights in the neurons of an older, smaller model as building blocks. Their machine-learning approach learns to expand the width and depth of the larger model in a data-driven way. Credit: Massachusetts Institute of Technology

It's no secret that OpenAI's ChatGPT has some incredible capabilities—for instance, the chatbot can write poetry that resembles Shakespearean sonnets or debug code for a computer program. These abilities are made possible by the massive machine-learning model that ChatGPT is built upon. Researchers have found that when these types of models become large enough, extraordinary capabilities emerge.

But bigger models also require more time and money to train. The training process involves showing hundreds of billions of examples to a model. Gathering so much data is an involved process in itself. Then come the monetary and environmental costs of running many powerful computers for days or weeks to train a model that may have billions of parameters.

"It's been estimated that training models at the scale of what ChatGPT is hypothesized to run on could take millions of dollars, just for a single training run. Can we improve the efficiency of these training methods, so we can still get good models in less time and for less money? We propose to do this by leveraging smaller language models that have previously been trained," says Yoon Kim, an assistant professor in MIT's Department of Electrical Engineering and Computer Science and a member of the Computer Science and Artificial Intelligence Laboratory (CSAIL).

Rather than discarding a previous version of a model, Kim and his collaborators use it as the building blocks for a new model. Using machine learning, their method learns to "grow" a larger model from a smaller model in a way that encodes knowledge the smaller model has already gained. This enables faster training of the larger model.

Their technique saves about 50% of the computational cost required to train a large model, compared to methods that train a new model from scratch. Plus, the models trained using the MIT method performed as

well as, or better than, models trained with other techniques that also use smaller models to enable faster training of larger models.

Reducing the time it takes to train huge models could help researchers make advancements faster with less expense, while also reducing the carbon emissions generated during the training process. It could also enable smaller research groups to work with these massive models, potentially opening the door to many new advances.

"As we look to democratize these types of technologies, making training faster and less expensive will become more important," says Kim, senior author of a paper on this technique.

Kim and his graduate student Lucas Torroba Hennigen wrote the paper with lead author Peihao Wang, a graduate student at the University of Texas at Austin, as well as others at the MIT-IBM Watson AI Lab and Columbia University. The research will be presented at the International Conference on Learning Representations, May 1–5.

The bigger the better

Large language models like GPT-3, which is at the core of ChatGPT, are built using a neural network architecture called a transformer. A neural network, loosely based on the human brain, is composed of layers of interconnected nodes, or "neurons." Each neuron contains parameters, which are variables learned during the training process that the neuron uses to process data.

Transformer architectures are unique because, as these types of neural network models get bigger, they achieve much better results.

"This has led to an arms race of companies trying to train larger and larger transformers on larger and larger datasets. More so than other

architectures, it seems that transformer networks get much better with scaling. We're just not exactly sure why this is the case," Kim says.

These models often have hundreds of millions or billions of learnable parameters. Training all these parameters from scratch is expensive, so researchers seek to accelerate the process.

One effective technique is known as model growth. Using the model growth method, researchers can increase the size of a transformer by copying neurons, or even entire layers of a previous version of the network, then stacking them on top. They can make a network wider by adding new neurons to a layer or make it deeper by adding additional layers of neurons.

In contrast to previous approaches for model growth, parameters associated with the new neurons in the expanded transformer are not just copies of the smaller network's parameters, Kim explains. Rather, they are learned combinations of the parameters of the smaller model.

Learning to grow

Kim and his collaborators use machine learning to learn a linear mapping of the parameters of the smaller model. This linear map is a mathematical operation that transforms a set of input values, in this case the smaller model's parameters, to a set of output values, in this case the parameters of the larger model.

Their method, which they call a learned Linear Growth Operator (LiGO), learns to expand the width and depth of larger network from the parameters of a smaller network in a data-driven way.

But the smaller model may actually be quite large—perhaps it has a hundred million parameters—and researchers might want to make a

model with a billion parameters. So the LiGO technique breaks the linear map into smaller pieces that a machine-learning algorithm can handle.

LiGO also expands width and depth simultaneously, which makes it more efficient than other methods. A user can tune how wide and deep they want the larger model to be when they input the smaller model and its parameters, Kim explains.

When they compared their technique to the process of training a new model from scratch, as well as to model-growth methods, it was faster than all the baselines. Their method saves about 50 percent of the computational costs required to train both vision and language models, while often improving performance.

The researchers also found they could use LiGO to accelerate transformer training even when they didn't have access to a smaller, pretrained model.

"I was surprised by how much better all the methods, including ours, did compared to the random initialization, train-from-scratch baselines." Kim says.

In the future, Kim and his collaborators are looking forward to applying LiGO to even larger models.

More information: Learning to Grow Pretrained Models for Efficient Transformer Training. openreview.net/pdf?id=cDYRS5iZ16f

This story is republished courtesy of MIT News (web.mit.edu/newsoffice/), a popular site that covers news about MIT research, innovation and teaching.