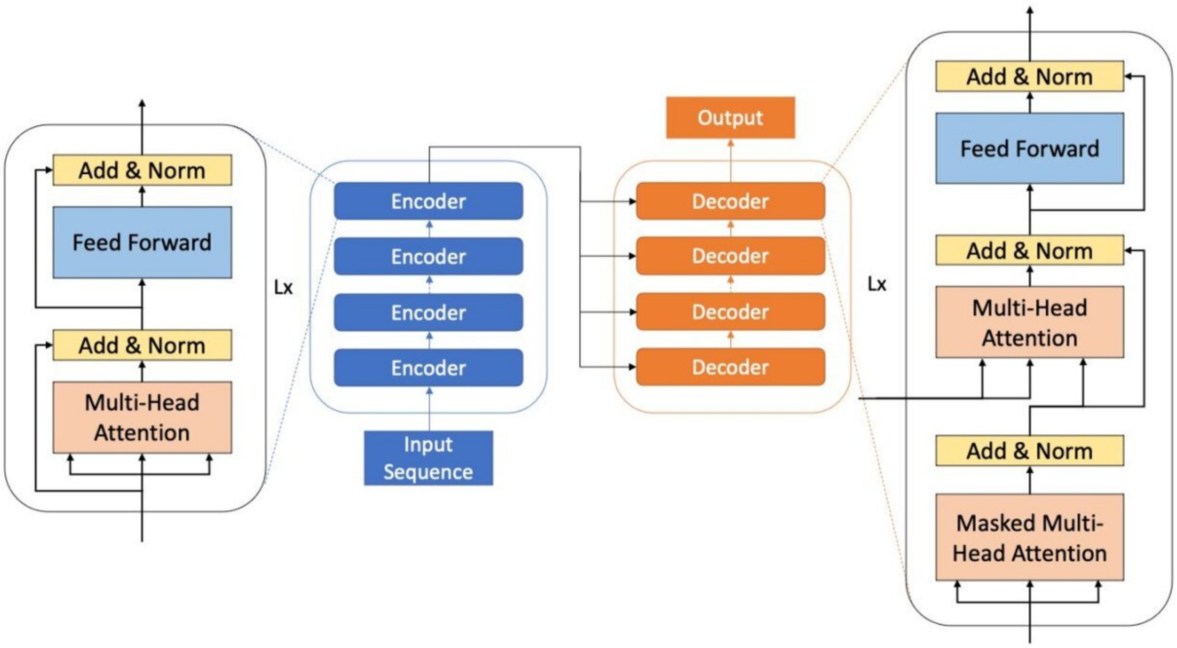
# Going big: World's fastest computer takes on large language modeling

May 14 2024, by Katie L Bethea



Transformer architecture. Credit: *arXiv* (2023). DOI: 10.48550/arxiv.2312.12705

A team led by researchers at the Department of Energy's Oak Ridge National Laboratory explored training strategies for one of the largest artificial intelligence models to date with help from the world's fastest supercomputer.

The findings could help guide training for a new generation of AI models for scientific research.

The study led by ORNL's Sajal Dash, Feiyi Wang and Prasanna Balaprakash employed Frontier, the world's first exascale supercomputer, to run the initial stages of training on a large language model similar to OpenAI's ChatGPT. The research team used a set of test data to project how models with 22 billion, 175 billion, and 1 trillion parameters, or variables, could run across 128 and later 384 of Frontier's more than 9,400 nodes. The team didn't attempt to train a full model to completion.

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

"This study and our findings aren't so much a manual as a potential set of guidelines for users training a large model," Dash said. "They can draw from our experience to decide how to use Frontier's resources to train their particular model and make the most effective use of their allotted computing time."

The team will present the study at the International Supercomputing Conference High Performance 2024 in May in Hamburg, Germany. Fellow scientists Isaac Lyngaas, Junqi Yin, Xiao Wang and Guojing Cong of ORNL and Romaine Egele of Paris-Saclay University also collaborated on the study.

The study focused less on model development than on pinpointing the most efficient ways to exploit the graphics processing units, or GPUs, that power Frontier and similar supercomputers and putting them to work training an AI. Each of Frontier's nodes relies on four AMD MI250X GPUs for a total of more than 75,000 GPUs.

The training ran for a few hours on about 100 million tokens—basic units of text such as words and characters—of test data. That's about a ten-thousandth of the necessary data to train a trillion-parameter model to completion and an even smaller fraction of the necessary time.

The research team used the data from those runs to calculate how a trillion-parameter model might perform if trained to completion on Frontier.

"This study was largely an exercise to show we can train this particular size of model on Frontier at this particular scale with this particular level of efficiency," Wang said. "We didn't get anywhere near the finish line of a complete large language model."

Large language models loosely mimic the human brain in their ability to learn and recognize patterns in words and numbers and to improve on that learning over time with additional training. The goal: design a model that can absorb and adjust the lessons learned on training data and apply that knowledge consistently and accurately to new, unfamiliar data and tasks.

The vast datasets and powerful processors needed for such training have remained mostly out of reach of scholars and in the possession of private companies, which tend to guard those resources as proprietary and set strict conditions for use. Those conditions typically limit research opportunities and don't allow results to be easily verified.

But leadership-class supercomputers like Frontier, which awards computing time to scientific researchers through the DOE's Innovative and Novel Computational Impact on Theory and Experiment program, could enable a new generation of AI models to be trained more quickly if scientists find the right approach.

"Traditionally, this process has relied on expert knowledge or on trial and error," said Balaprakash, ORNL's director of AI programs. "One of the highlights of our work in this study is the automation of identifying high-performing strategies among a vast array of options. We leveraged DeepHyper, an open-source scalable tuning software, to automatically determine the optimal settings.

"We plan to extend this automated approach to fine-tune system-level performance and enhance efficiency at an extreme scale. Furthermore, we have democratized our methodologies and software for the benefit of the scientific community. This strategy ensures that our insights are widely accessible for future research on training large AI foundation models in science."

The larger the model and its training datasets, the better its performance—but also the higher its demand for computational power. Training a trillion-parameter large language model from the initial stages to completion without optimizations would take months even at Frontier's world-leading speeds.

The ORNL study examined approaches to data parallelism—a process used by supercomputers like Frontier to break a large problem into smaller problems to reach a solution more quickly—to train AI and how to port that training across proprietary frameworks of GPUs made by competing vendors.

"It's about finding the best combination of training strategies while getting the best throughput," Dash said. "Most deep-learning frameworks target the GPUs made by NVIDIA rather than the GPUs made by AMD that power Frontier. We wanted to see if existing models could run on Frontier, how to make the best use of Frontier's computing power and how to make that level of performance possible across GPU platforms.

"We can't train a model this size on a single GPU or a single node, for example, and every time we cross the barrier between nodes that requires more communication that consumes more time. How do we slice up the model across GPUs so that we can fit and train the model without losing too much time and energy communicating between nodes?"

The researchers found a blend of parallelism strategies worked best when tailored to the computing platform but said their work's far from finished.

"The efficiency we achieved on Frontier with this model was decent but not decent enough," Wang said. "At extreme scale, we achieved 30% efficiency—which means we left about 70% of Frontier's computing power on the floor. We need much more optimization to make the machine more efficient at this scale."

The team's next steps include training a model further with peer-reviewed scientific data across more nodes.

**More information:** Sajal Dash et al, Optimizing Distributed Training on Frontier for Large Language Models, *arXiv* (2023). DOI: 10.48550/arxiv.2312.12705

Provided by Oak Ridge National Laboratory

Citation: Going big: World's fastest computer takes on large language modeling (2024, May 14) retrieved 17 July 2024 from https://techxplore.com/news/2024-05-big-world-fastest-large-language.html