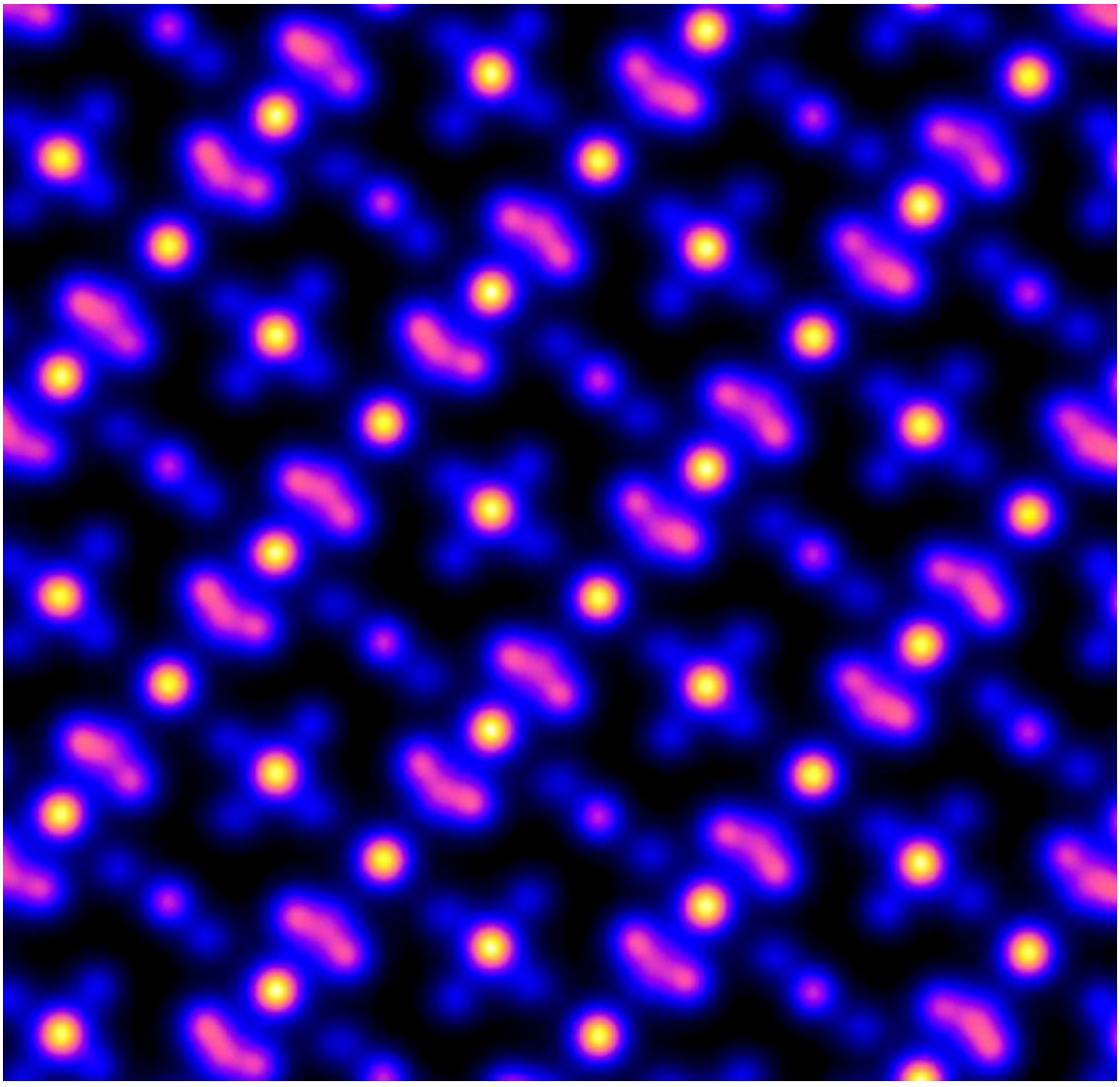
New deep learning techniques lead to materials imaging breakthrough

April 27 2022, by Elizabeth Rosenthal



The team's techniques dramatically increased the number of images that can be processed at once while training DNNs. Pictured here is one of many images of scanning transmission electron microscope data included in these scaling

efficiency-focused simulations. Credit: Junqi Yin/ORNL, U.S. Dept. of Energy

Supercomputers help researchers study the causes and effects—usually in that order—of complex phenomena. However, scientists occasionally need to deduce the origins of scientific phenomena based on observable results. These so-called inverse problems are notoriously difficult to solve, especially when the amount of data that must be analyzed outgrows traditional machine-learning tools.

To better understand inverse problems, a team from the US Department of Energy's (DOE's) Oak Ridge National Laboratory (ORNL), NVIDIA, and Uber Technologies developed and demonstrated two new techniques within a widely used communication library called Horovod.

Developed by Uber, this platform trains deep neural networks (DNNs) that use algorithms to imitate and harness the decision-making power of the human brain for scientific applications. Because Horovod relies on a single coordinator to provide instructions to many different workers (i.e., GPUs in this case) to complete this process, large-scale deep-learning applications often encounter significant slowdowns during training.

The researchers' methods removed repetitive steps from the traditional coordinator-worker process to increase the speed and outperform existing approaches, thereby enabling them to uncover the first-ever approximate solution to an age-old inverse problem in the field of materials imaging. Their results were published in the Proceedings of the 19th USENIX Symposium on Networked Systems Design and Implementation.

"As far as we know, this is the most floating-point operations per second ever achieved for the distributed training of a convolutional neural

network," said Junqi Yin, a computational scientist in ORNL's Analytics and AI methods at Scale group. "We plan to use the resulting code, STEMDL, for benchmarking future generations of supercomputers."

Stronger together

To reduce coordinator-worker communication, which often involves repeating the same requests multiple times, the team introduced a response cache that stores the metadata from each request in Horovod. The first of the scientists' new strategies was this caching approach, which allows Horovod to immediately recognize and automatically calculate familiar requests without delaying DNN training.

Their second new technique involves grouping the mathematical operations of multiple DNN models, which streamlines tasks and improves scaling efficiency—the total number of images processed per training step—by taking advantage of the similarities in each model's calculations. This process leads to significant improvements in power usage as well.

By strategically grouping these models, the team aims to eventually train a single model on multiple GPUs and achieve the same efficiency obtained when training one model per GPU.

Josh Romero, a developer technology engineer at NVIDIA, incorporated the new tactics into Horovod to enable users to train DNNs more efficiently on high-performance computing machines of any size.

"All workers must agree on the order of operations and on what information is going to be distributed at any given time," Romero said. "We found a way to improve this logistical process."

Both methods enhanced Horovod's performance individually, but

combining them nearly doubled scaling efficiency, which the team measured by running the STEMDL code on all 27,600 GPUs of the IBM AC922 Summit system. Summit, the nation's fastest supercomputer, is located at ORNL's Oak Ridge Leadership Computing Facility, a DOE Office of Science user facility.

"These capabilities are what allowed us to train a single neural network distributed across all of Summit with much higher scaling efficiency and much better computing performance than was previously possible at large scales," said Nouamane Laanait, former computational scientist at ORNL and principal investigator of the team's Summit allocation, which was granted through the Innovative and Novel Computational Impact on Theory and Experiment program.

Convolutional neural networks such as STEMDL are ideal DNNs for image analyses. The team designed this application specifically to solve a long-standing materials-imaging inverse problem, which requires precise analysis of scanning transmission electron microscope data.

"One of the advantages of using neural network models is that you can incorporate a lot of factors that are difficult to encode in mathematical approaches to solving inverse problems," Laanait said. "By training these models on datasets, you can teach them to overlook noise and other imperfections."

All-encompassing architecture

Summit's unique components made this research possible. For example, distributing DNN training among the supercomputer's GPUs revealed the performance bottlenecks present in traditional Horovod calculations. These roadblocks accumulate throughout the training process before they become apparent as they begin to hamper compute times, which makes them difficult or impossible to see on smaller systems.

"Within an hour you know how precise the solution is, which allows you to tweak the prototype much faster than on smaller systems, which can take days or weeks to determine how exact a model is or how well you mapped the problem to your model," Laanait said.

Additionally, Summit has high-bandwidth communication pathways to move data from place to place, and its local storage system—known as the burst buffer—has sufficient memory to allow researchers to simulate and store more than a terabyte of data on each node. Finally, the NVIDIA Tensor Cores—specialized processing units ideal for deep learning applications—sped up the team's code and helped them reach higher performance levels than would have been possible on traditional processors.

The team's findings could be applied to existing deep learning applications and previously unsolved inverse problems to answer fundamental science questions. Going forward, the researchers hope to recreate their results using less compute power and train even larger models required by the ever-increasing amount of data generated by experimental facilities.

"Combining larger datasets and models with more compute power usually increases the effectiveness of DNNs," Laanait said. "We don't know what the ceiling is for these improvements, so the only way to find out is by continuing to experiment."

More information: Joshua Romero et al, Accelerating Collective Communication in Data Parallel Training across Deep Learning Frameworks, Proceedings of the 19th USENIX Symposium on Networked Systems Design and Implementation (April 4–6, 2022). www.usenix.org/system/files/ns … i22-paper-romero.pdf

Provided by Oak Ridge National Laboratory