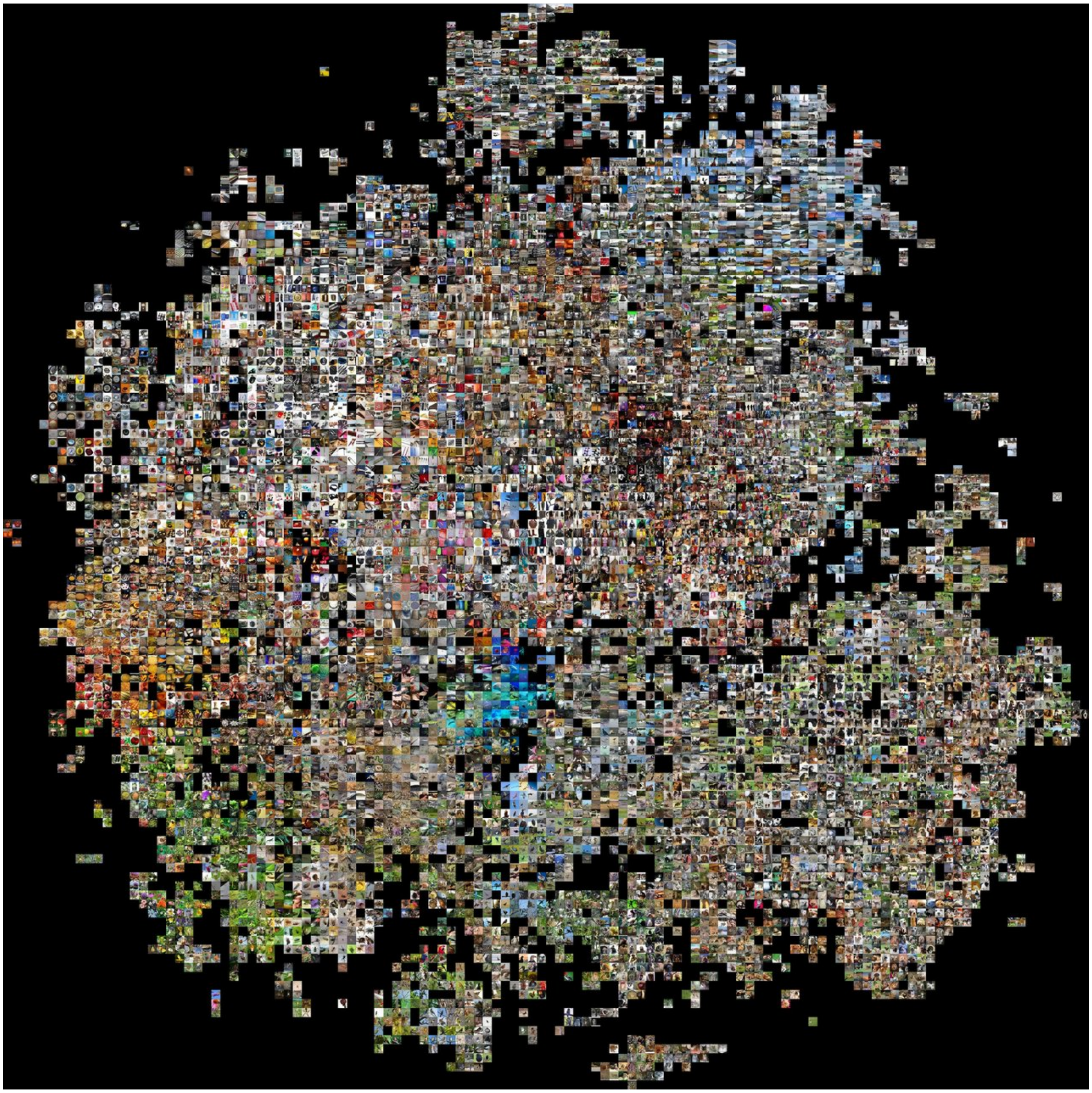
Supercomputing speeds up deep learning training

November 13 2017



Two-dimensional embedding of images from the ImageNet database, extracted by a convolutional neural network using Caffe. Credit: Andrej Karpathy

A team of researchers from the University of California, Berkeley, the University of California, Davis and the Texas Advanced Computing Center (TACC) published the results of an effort to harness the power of supercomputers to train a deep neural network (DNN) for image recognition at rapid speed.

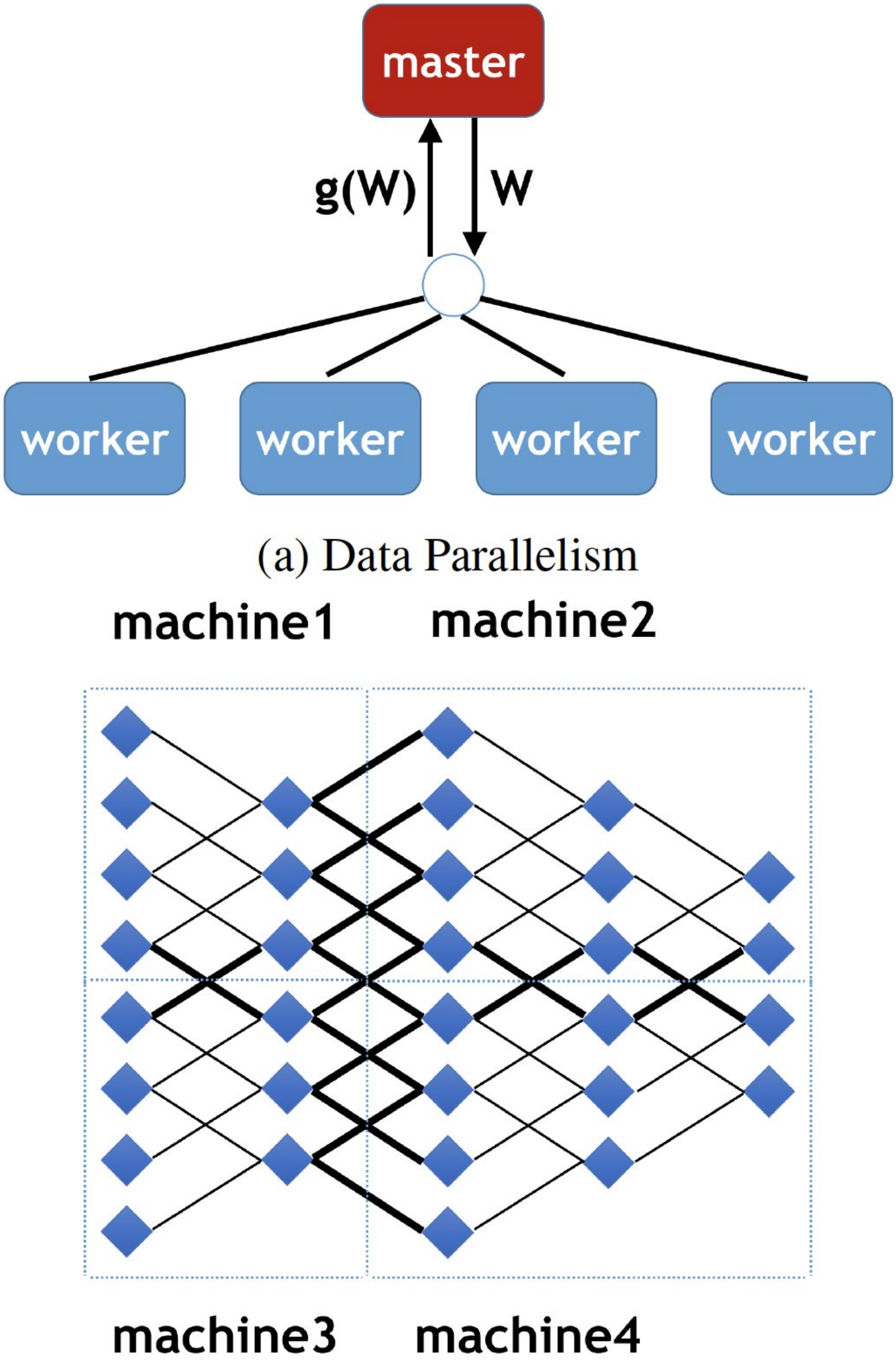
The researchers efficiently used 1024 Skylake processors on the Stampede2 supercomputer at TACC to complete a 100-epoch ImageNet training with AlexNet in 11 minutes - the fastest time recorded to date. Using 1600 Skylake processors they also bested Facebook's prior results by finishing a 90-epoch ImageNet training with ResNet-50 in 32 minutes and, for batch sizes above 20,000, their accuracy was much higher than Facebook's. (In recent years, the ImageNet benchmark—a visual database designed for use in image recognition research—has played a significant role in assessing different approaches to DNN training.)

Using 512 Intel Xeon Phi chips on Stampede2 they finished the 100-epoch AlexNet in 24 minutes and 90-epoch ResNet-50 in 60 minutes.

"These results show the potential of using advanced computing resources, like those at TACC, along with large mini-batch enabling algorithms, to train deep neural networks interactively and in a distributed way," said Zhao Zhang, a research scientist at TACC, a leading supercomputing center. "Given our large user base and huge capacity, this will have a major impact on science."

They published their results in *Arxiv* in November 2017.

The DNN training system achieved state-of-the-art "top-1" test accuracy, which means the percentage of cases where the model answer (the one with highest probability) is exactly the expected answer. Using ResNet-50 (a Convolutional Neural Networks developed by Microsoft that won the 2015 ImageNet Large Scale Visual Recognition Competition and surpasses human performance on the ImageNet dataset) they achieved an accuracy of more than 75 percent - on par with Facebook and Amazon's batch training levels. Scaling to the batch size of the data 32,000 in this work only lost 0.6 percent top-1 accuracy.



(a) Data Parallelism

(b) Model Parallelism

Schematic showing data parallelism vs. model parallelism, as they relate to neural network training. Credit: Yang You, Zhao Zhang, Cho-Jui Hsieh, James Demmel, Kurt Keutzer

Currently deep learning researchers need to use trial-and-error to design new models. This means they need to run the training process tens or even hundreds of times to build a model.

The relatively slow speed of training impacts the speed of science, and the kind of science that researchers are willing to explore. Researchers at Google have noted that if it takes one to four days to train a neural network, this is seen by researchers as tolerable. If it takes one to four weeks, the method will be utilized for only high value experiments. And if it requires more than one month, scientists won't even try. If researchers could finish the training process during a coffee break, it would significantly improve their productivity.

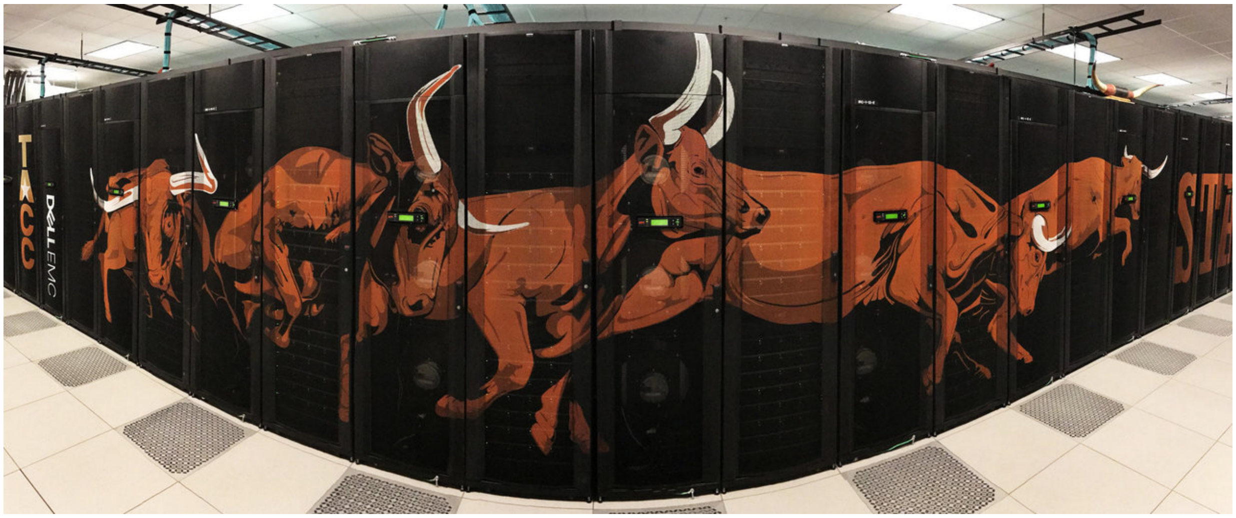
The group's breakthrough involved the development of the Layer-Wise Adaptive Rate Scaling (LARS) algorithm that is capable of distributing data efficiently to many processors to compute simultaneously using a larger-than-ever batch size (up to 32,000 items).

LARS incorporates many more training examples in one forward/backward pass and adaptively adjusts the learning rate between each layer of the neural network depending on a metric gleaned from the previous iteration.

As a consequence of these changes they were able to take advantage of the large number of Skylake and Intel Xeon Phi processors available on

Stampede2 while preserving accuracy, which was not the case with previous large-batch methods.

"For deep learning applications, larger datasets and bigger models lead to significant improvements in accuracy, but at the cost of longer training times," said James Demmel, "A professor of Mathematics and Computer Science at UC Berkeley. "Using the LARS algorithm, jointly developed by Y. You with B. Ginsburg and I. Gitman during an NVIDIA internship, enabled us to maintain accuracy even at a batch size of 32K. This large batch size enables us to use distributed systems efficiently and to finish the ImageNet training with AlexNet in 11 minutes on 1024 Skylake processors, a significant improvement over prior results."



Researchers used Stampede2 to train a deep neutral network on the ImageNet-1k benchmark set in minutes. Credit: Sean Cunningham, Texas Advanced Computing Center

The findings show an alternative to the trend of using specialized hardware - either GPUs, Tensor Flow chips, FPGAs or other emerging architectures—for deep learning. The team wrote the code based on Caffe and utilized Intel-Caffe, which supports multi-node training.

The training phase of a deep neural network is typically the most time-intensive part of deep learning. Until recently, the process accomplished by the UC Berkeley-led team would have taken hours or days. The advances in fast, distributed training will impact the speed of science, as well as the kind of science that researchers can explore with these new methods.

The experiment is part of a broader effort at TACC to test the applicability of CPU hardware for deep learning and machine learning applications and frameworks, including Caffe, MXNet and TensorFlow.

TACC's experts showed how they when scaling Caffe to 1024 Skylake processors using resNet-50 processors, the framework ran with about 73 percent efficiency—or almost 750 times faster than on a single Skylake processor.

"Using commodity HPC servers to rapidly train deep learning algorithms on massive datasets is a powerful new tool for both measured and simulated research," said Niall Gaffney, TACC Director of Data Intensive Computing. "By not having to migrate large datasets between specialized hardware systems, the time to data driven discovery is reduced and overall efficiency can be significantly increased."

As researchers and scientific disciplines increasingly use machine and deep learning to extract insights from large scale experimental and simulated datasets, having systems that can handle this workload are important.

Recent results suggest such systems are now available to the openscience community through national advanced computing resources like Stampede2.

More information: ImageNet Training in Minutes, arXiv:1709.05011 [cs.CV] arxiv.org/abs/1709.05011

Provided by University of Texas at Austin

Citation: Supercomputing speeds up deep learning training (2017, November 13) retrieved 19 April 2024 from https://techxplore.com/news/2017-11-supercomputing-deep.html