Improve machine learning performance by dropping the zeros

August 23 2021

Credit: Pixabay/CC0 Public Domain

KAUST researchers have found a way to significantly increase the speed of training. Large machine learning models can be trained significantly faster by observing how frequently zero results are produced in distributed machine learning that use large training datasets.

AI models develop their "intelligence" by being trained on datasets that have been labeled to tell the model how to differentiate between different inputs and then respond accordingly. The more labeled data that goes in, the better the model becomes at performing whatever task it has been assigned to do. For complex deep learning applications, such as self-driving vehicles, this requires enormous input datasets and very long training times, even when using powerful and expensive highly parallel supercomputing platforms.

During training, small learning tasks are assigned to tens or hundreds of computing nodes, which then share their results over a communications network before running the next task. One of the biggest sources of computing overhead in such parallel computing tasks is actually this communication among computing nodes at each model step.

"Communication is a major performance bottleneck in distributed deep learning," explains Jiawei Fei from the KAUST team. "Along with the fast-paced increase in model size, we also see an increase in the proportion of zero values that are produced during the learning process, which we call sparsity. Our idea was to exploit this sparsity to maximize effective bandwidth usage by sending only non-zero data blocks."

Building on an earlier KAUST development called SwitchML, which optimized internode communications by running efficient aggregation code on the network switches that process data transfer, Fei, Marco Canini and their colleagues went a step further by identifying zero results and developing a way to drop transmission without interrupting the synchronization of the parallel computing process.

"Exactly how to exploit sparsity to accelerate distributed training is a challenging problem," says Fei. "All nodes need to process data blocks at the same location in a time slot, so we have to coordinate the nodes to ensure that only data blocks in the same location are aggregated. To overcome this, we created an aggregator process to coordinate the workers, instructing them on which block to send next."

The team demonstrated their OmniReduce scheme on a testbed consisting of an array of graphics processing units (GPU) and achieved an eight-fold speed-up for typical deep learning tasks.

"We are now adapting OmniReduce to run on programmable switches using in-network computation to further improve performance," Fei says.

More information: Jiawei Fei et al, Efficient sparse collective communication and its application to accelerate distributed deep learning, *Proceedings of the 2021 ACM SIGCOMM 2021 Conference* (2021). DOI: 10.1145/3452296.3472904

Provided by King Abdullah University of Science and Technology

Citation: Improve machine learning performance by dropping the zeros (2021, August 23) retrieved 26 April 2024 from https://techxplore.com/news/2021-08-machine-zeros.html