Divya Koya

CS 598 AI Efficiency for Systems and Algorithms

Professor Minjia Zhang

November 20, 2024

Review 21: Neo + ZionEx

This paper aims to address the issues that come with training large Deep Learning Recommendation Models (DLRMs) and finding ways to optimize their training. DLRMs are integral to many companies today—for example, Netflix can use them for movie recommendations, NY Times could use it for news feed organization, and Instagram could use them for personalized advertisements. DLRMs differ from traditional Deep Neural Networks (DNNs) in that they not only have compute-intensive operations, but also data-intensive embedding operators as well. The parallelism methods that DNNs use cannot be applied to DLRMs, so there is a need for a high-performance synchronous training solution with memory-efficient computation that scales while preserving the accuracy of the model.

As such, the authors propose Neo, a system that addresses the need for optimized DLRM training and has been deployed into production as well. Neo leverages 4 major embedding optimizations. The first is hybrid kernel fusion, which reduces the overhead of launching thousands of CUDA kernels by fusing embedding lookups on every GPU into one CUDA kernel, and by fusing the backward pass with the sparse optimizer to do gradient computations and updates in one CUDA kernel as well. The second is the use of 4D parallelism—table, row, column, and data parallelism—to efficiently partition and store embedding tables on different nodes in order to optimize embedding operators. The third is a multi-level memory hierarchy that makes use of HBM, DRAM, and SSDs to speed up data accesses, along with row-wise sparse optimizers, mixed-precision training, and advanced factorization techniques to further compress the amount of memory being used. And lastly, Neo uses ZionEx, a hardware system co-designed with Neo to optimize inter-node communications for DLRM training.

I believe that the greatest strengths of this work lie in its significant improvements upon the state-of-the-art and its ease of use. The paper evaluates Neo on 128 GPUs using 16 ZionEx nodes, and it shows that Neo brings a near-40x improvement on existing systems that train 12-trillion-parameter DLRM models. Furthermore, Neo has already been deployed into production and is already in use, which implies that it is not only easy to use but is also able to handle production-level systems.

A further avenue for research would be to look into further improving the communication networks of this system, because as DLRM models get bigger and bigger, the communication between nodes will start to become the bottleneck on the training time of these models. It might be interesting to look into using quantization or other compression methods, or possibly using

multi-path data routing or programmable network switches that route data through optimal network paths in datacenters.