Accelerating Distributed MoE Training and Inference with Lina

Jiamin Li∗, Yimin Jiang‡, Yibo Zhu, Cong Wang∗, Hong Xu†
∗City University of Hong Kong, ‡ByteDance Inc., †The Chinese University of Hong Kong

Introduction
Mixture-of-Experts (MoE): a popular way to curb the computation cost of deep learning models.

[MoE in language models] MoE layer replaces the FFN layer in Transformer. It consists of multiple FFNs as experts, and a gating network. The gating network dispatches the token to a small number of experts (top-1, top-2).

[Distributed MoE] Data parallelism and expert parallelism are applied. It allocates one unique GPU for each expert and uses all-to-all to exchange tokens.

Motivation
Why is all-to-all the bottleneck in distributed MoE?

[1. Synchronous all-to-all with large data transfer]
Stream a | Stream b | Stream c
---|---|---
Gate | All-to-all | All-to-all | Combine
0 | 3.1 | 10.7 | 12.1 | 18.7 | 20.3 | ms

74.9% of the running time of one MoE layer.

Training
Prolonged all-to-all with allreduce] In the backward pass, all-to-all and allreduce control their own process group and overlap, they contend for the network bandwidth and their completion times are severely prolonged.

Inference
Skewed expert popularity] The token-to-expert distribution in inference is purely workload-driven. Expert popularity is highly skewed in sharp contrast to training.

Design
Lina replicates popular experts on proportionally more devices to balance the workload.

How to know the expert popularity a priori?

[Pattern in expert selection] Tokens that have selected the same expert in layer i tend to select the same expert again in layer i + 1.

[Two-phase scheduling] a. Resource scheduling based on estimated popularity
   Estimate with patterns profiled during training
b. Low-overhead fine-tuning on actual routing decision

Evaluation
Lina reduces the 95%ile inference time by an average of 1.63x

Lina reduces the training step time by up to 1.73x.