Supplement for: TA-MoE: Topology-Aware Large Scale Mixture-of-Expert Training

Anonymous Author(s)
Affiliation
Address
email

1 Appendix

1.1 Perplexity Evaluation Results

Table 1: **Perplexity Evaluation Result.** To further validate the model convergence performance, we list the perplexity (PPL) at 10w step (near 7 days) of GPT-Medium (12 layers, hidden size 1024, intermediate size 2048, GShard, Capacity factor 1.2) with different expert numbers on the openwebtext2 dataset. Combined with the information in Figure 3 of the paper, we can find that TA-MoE does not influence the convergence of the model when compared with the well-known FastMoE baseline.

Expert Scale	TA-MoE Valid PPL	Baseline Valid PPL
8	17.97	18.12
16	15.18	15.39
32	13.37	13.53
48	12.55	12.49

1.2 Data Dispatch Distribution

Figure 1 further elaborates the data dispatch patterns of GPU rank 0-7 (within a node) on 8, 16, 32, 48 experts. We also list the Rank-0 dispatch information of FastMoE with even distribution method as the baseline. On a single node topology, each rank prefers to dispatch data to the expert within a node. The topology influence on the data dispatch preference is relatively small, because the bandwidth variance within a node is small. Multi-node topology results show a consistent "ladder-like" distribution trend that the ranks within a node has a high preference to dispatch the data to intra-node rank group, instead of transferring data to inter-node ranks. These results verify the effectiveness of the proposed adaptive topology-aware method.

1.3 Tests on Swin Tranformer Based MoE Tasks

To further validate the generality of TA-MoE, we also carry out an experiment of Vision Tasks on ImageNet-1k dataset. The well-known vision transformer architecture Swin Transformer [1] is picked as the base model. The detailed model configurations are listed in Table 2. We evaluate the speedup on Cluster A with 16 and 32 GPUs as an illustration, where 16 GPUs configurations represents symmetric tree topology and 32 GPUs configurations represents asymmetric tree topology. As shown in Figure 2, we can achieve 1.18x and 1.20x speedup when compared with FastMoE on 16, and 32 GPUs, respectively.

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

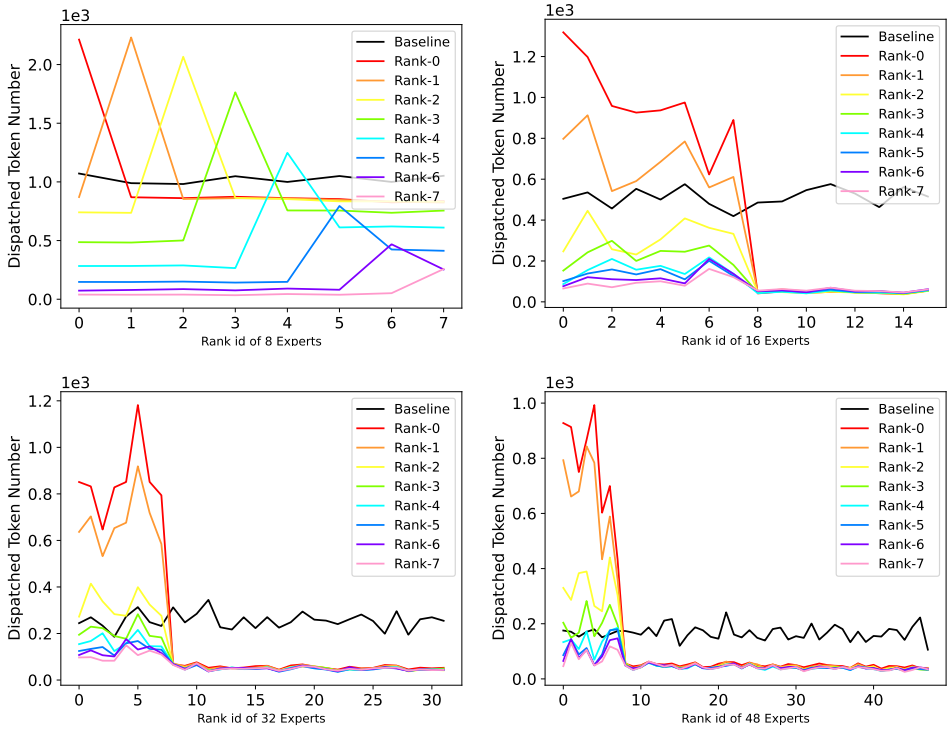

Figure 1: The data dispatch distribution of rank 0 to 7 on 16, 32, 48 GPUs and experts.

Table 2: Detailed specifications of the Swin-Transformer model.

Name	Layers	Gate	Stage 1	Stage 2	Stage 3	Stage 4	Capacity factor
Swin Transformer v1	12	GShard	concat 4x4, 96-d, LN {win.sz. 7x7, dim 96, head3} x2	concat 2x2, 192-d, LN {win.sz. 7x7, dim 192, head6} x2	concat 2x2, 384-d, LN {win.sz. 7x7, dim 384, head12} x6	concat 2x2, 768-d, LN {win.sz. 7x7, dim 768, head324} x2	1.2

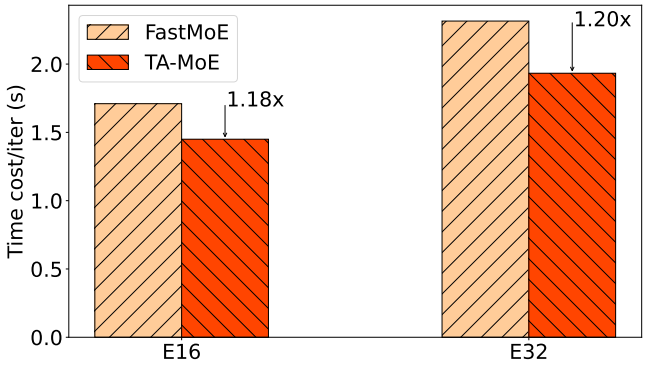

Figure 2: Speedup of TA-MoE over FastMoE on Swin Transformer Based Model.

References

[1] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. *CoRR*, abs/2103.14030, 2021.