Journal Club

By Malika Bakhtawar

2024/07/12

Research Article Information

Paper title

“Accelerating Heterogeneous Tensor Parallelism via Flexible Workload Control”

Published Date

21 Jan, 2024

Authors

Wang, Zhigang, Xu Zhang, Ning Wang, Chuanfei Xu, Jie Nie, Zhiqiang Wei, Yu Gu, and Ge Yu.

Paper Highlights

- Focuses on tensor parallelism within heterogeneous computing environments.
- Introduces novel solutions to mitigate the impact of stragglers during parallel model training.
- Addresses the challenge of slower processing devices affecting overall training efficiency.
- Proposes methods to optimize and balance workloads for faster and more efficient training.

Background

- Many Big IT companies are building AI compute clusters equipped with homogenous accelerators such as AI Research SuperCluster RSC in Meta.
- However due to heterogeneity parallel tasks usually proceed at very different speeds.
- Techniques such as Data Migration and customizing input data batch size exits but they are for traditional data parallelism.

Major Contributions

- They proposed ZERO-resizing, a novel matrix resizing approach to dynamically balance workloads under tensor parallelism.
- They introduced another lightweight migration approach to cope with heavily heterogeneous environments.
- They further build hybrid solution SEMI-migration on top of the two approaches.
- Finally they performed extensive experimental studies using a well known foundation model ViT with billions of parameters.

ZERO-RESIZING: BALANCING WORKLOADS BY RESIZING MATRICES

- It temporary resizes the matrices involved in tensor computation
- Immediately prunes some columns or rows (dual) in matrices once a straggling phenomena happens to reduce the workload.
- Pruning is performed in both forward- and backward-propagations.
- To recover the dimensions in backward propagation they impute the missing data to avoid abnormally terminated propagation.

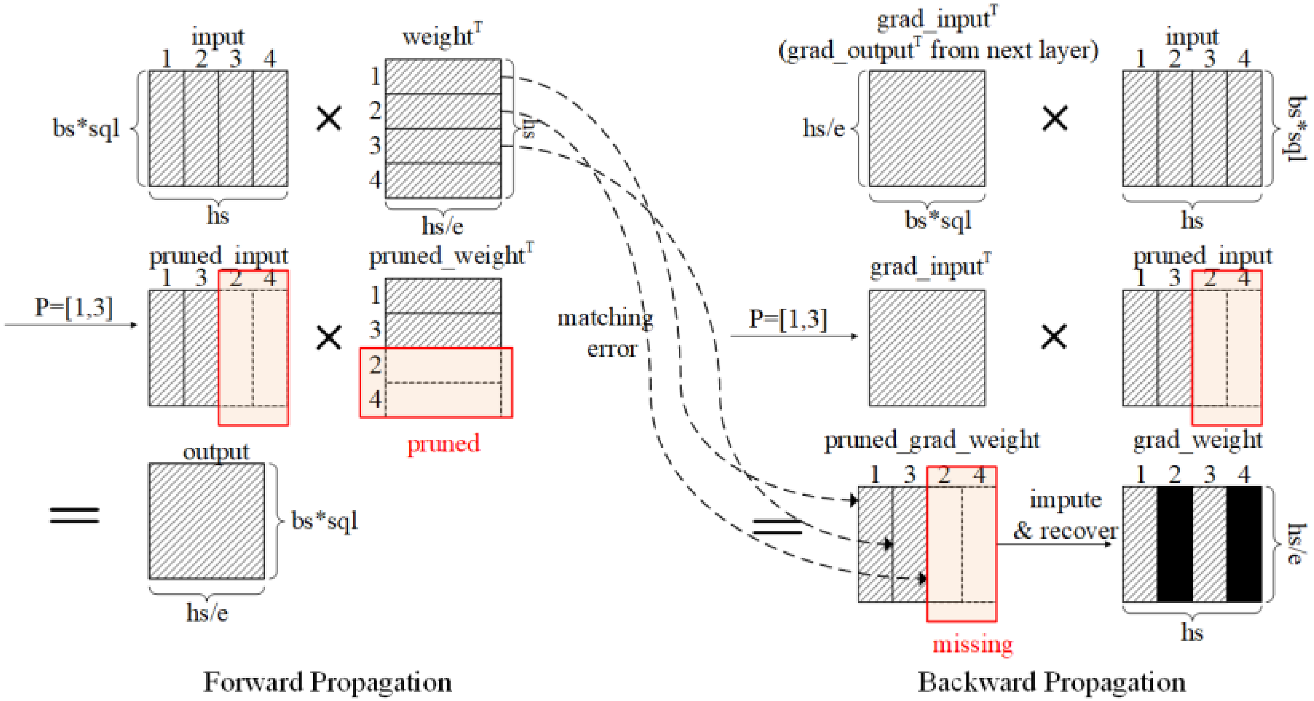


Fig. 2. Matrix pruning and imputation process ($\gamma = 0.5$)

ZERO-RESIZING: BALANCING WORKLOADS BY RESIZING MATRICES

Advantages

It adaptively adjust its structure without the need for migrating data across different computational units or devices. This can significantly reduce additional time associated with data movement.

It enables the model to dynamically react to changes in heterogeneity among devices like how many columns/rows and which column/row should be pruned.

Disadvantage.

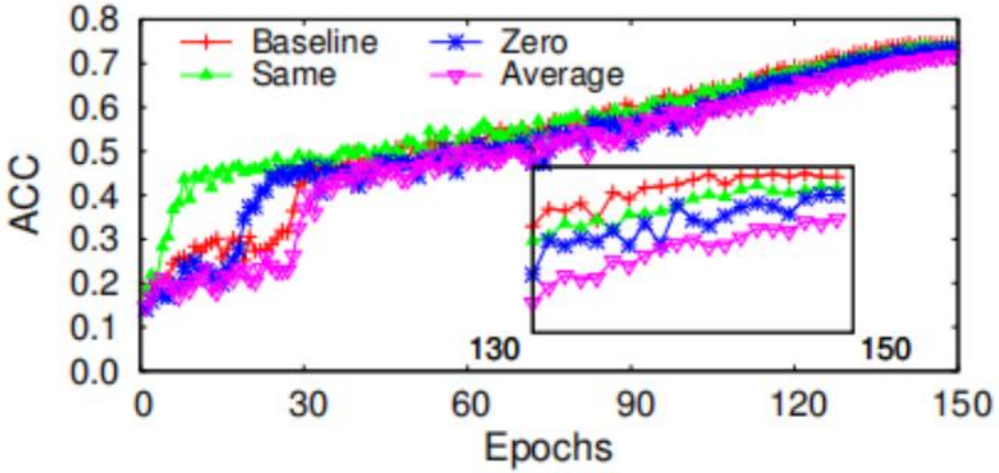
While zero resizing mitigates migration costs and adapts to heterogeneity, it inevitably leads to model accuracy degradation when pruning occurs.

Impact of different imputation policies on the model accuracy

- They maintain a lookup table with a series of three tuples <layer_name, matrix_name, P> , where P is set of indexes related to pruned columns.
- But the issue is what value should be imputed into missing dimensions.

They tested three policies: (γ=0.5)

- Same (use same values calculated in previous iteration)
- Average (average from unpruned dimensions)
- Zero (the uniform zero value)



The final choice is Zero which balances space complexity and accuracy.

Lightweight Workload Migration

- Once a straggler is detected same in the resizing technique.
- Gamma (γ) is determined , this time it will act as migration ratio that calculate how many columns should be migrated.
- After determining the data is evenly distributed across other fast devices.

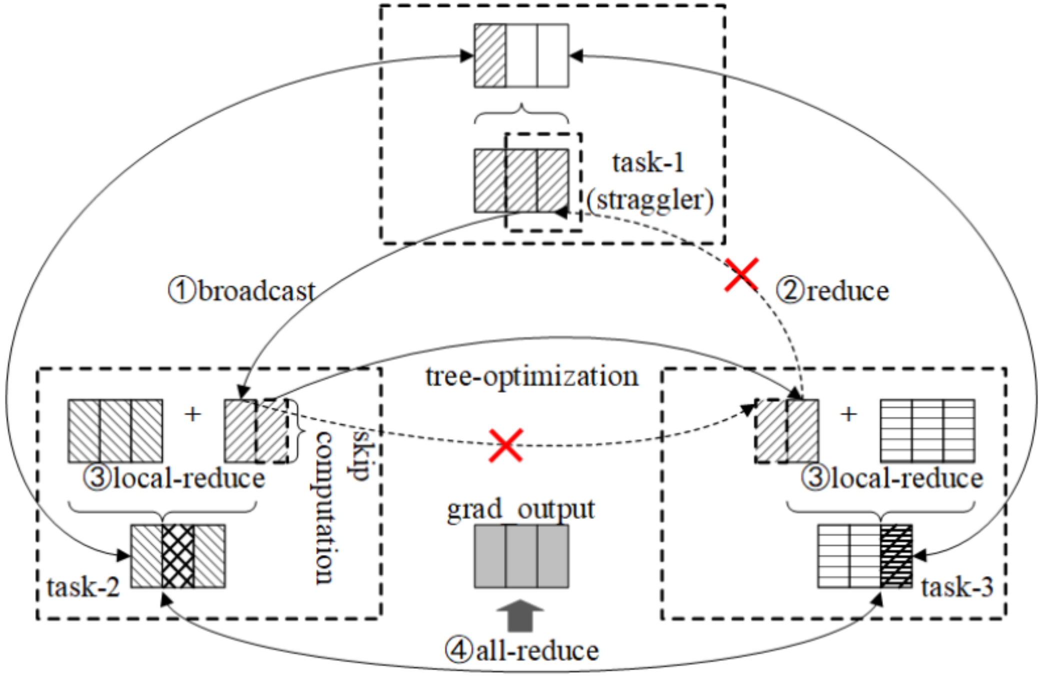


Fig. 4. Illustration of the sending-collecting migration

SEMI-MIGRATION: A HYBRID BALANCING SOLUTION

- It solely runs resizing if the speed difference is tolerable; and otherwise re-assigns workloads via migration.

How it Works

- When it encountered many stragglers it first sorts them in descending order with respect to their speed differences.
- Then the top-stragglers used a migration solution, while others employ resizing solution.
- It balances the trade off between model accuracy and efficiency of the training.

Results and Discussions

They also tested two types of Pruning

- Random pruning
- Priority Pruning

❖ Pri selectively pruned those certain weight dimensions that are less important having small variations in their gradient updates.

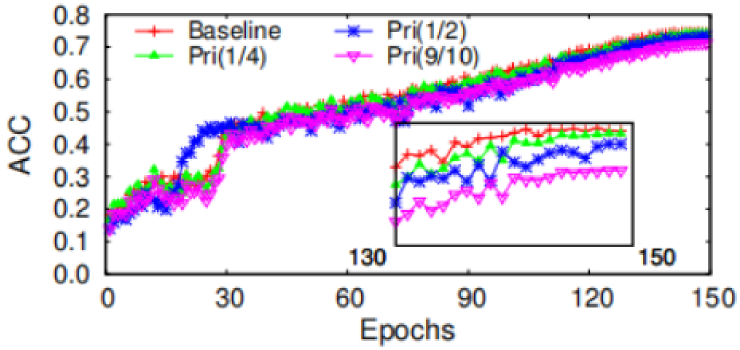


Fig. 7. Accuracy variation with fixed straggling skewness (ViT-1B)

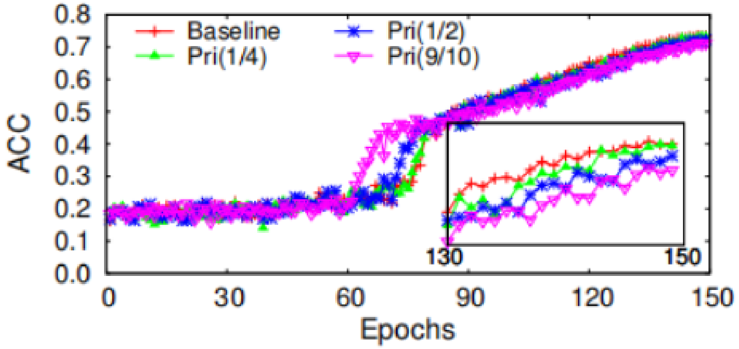


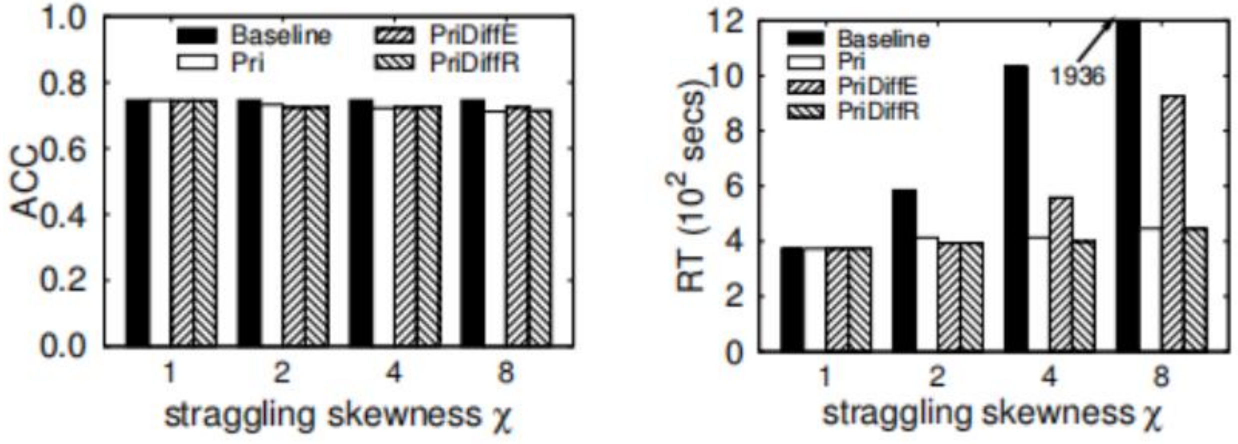
Fig. 8. Accuracy variation with fixed straggling skewness (ViT-3B)

Results and Discussions

- **Straggling Skewness χ:**

χ shows the number of stragglers.

- **RT (Runtime)**: An averaged elapsed time of an epoch.



(a) Model accuracy

(b) Training efficiency

Fig. 9. Overall performance in heterogeneous environments (ViT-1B)

Conclusion

- This paper studies the straggling issue of tensor parallelism when training foundation models in complex heterogeneous environments.
- To overcome these issues they proposed three solutions to handle the stragglers effectively such as ZERO-Resizing, Light weight migration and SEMI Migration.
- They achieved improved performance in dynamically balance workloads to reduce waiting costs, in terms of efficiency and accuracy.

Thank You!