Peng LIANG, Linbo QIAO, Yanqi SHI, Hao ZHENG, Yu TANG, Dongsheng LI, 2025. Memory-efficient tensor parallelism for long-sequence Transformer training. *Frontiers of Information Technology & Electronic Engineering*, 26(5):770-787. https://doi.org/10.1631/FITEE.2400602

Memory-efficient tensor parallelism for long-sequence Transformer training

Key words: Distributed learning; Large language model (LLM); Long sequence; Machine learning system; Memory efficiency; Tensor parallelism

Dongsheng LI
E-mail: dsli@nudt.edu.cn
ORCID: https://orcid.org/0000-0001-9743-2034

Motivation

1. Training language models with long sequences is crucial for handling long-context tasks. However, it requires $O(s^2)$ memory for a sequence with s tokens to complete the attention computation in Transformer models.

2. To train models with long sequences, up to now, several parallelism methods have been proposed to reduce the memory footprint. However, current parallelism methods could only reduce memory footprint to $\Omega(1/\#\text{device})$.

3. FlashAttention employs tiling techniques to optimize the attention computation process, and reduces the memory footprint of computing attention from $O(s^2)$ to $O(s)$. However, FlashAttention is a work designed for a single device.

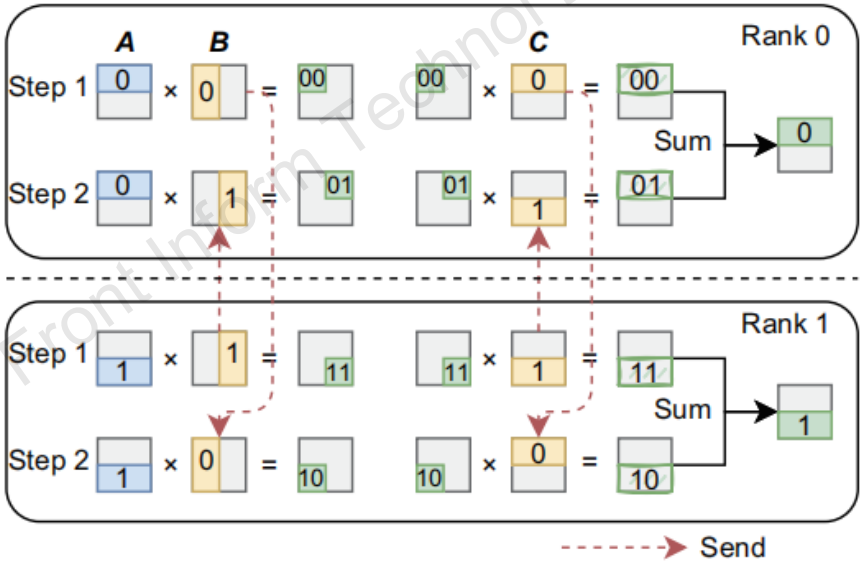
Main idea

1. With the inspiration of FlashAttention, the proposed work METP (memory-efficient tensor parallelism) decomposes the computation tasks by tiling to handle the large memory overhead problem of training LLMs with long sequences.

2. METP applies a two-level loop algorithm that decomposes the attention computation into p^3 tasks (p is the parallel degree), and overlaps the computation with communication to improve throughput.

Method

1. METP decomposes the computation of $\boldsymbol{O}=f(\boldsymbol{AB})\boldsymbol{C}$ into p^2 tasks and each device executes p of them. The intermediate result $\boldsymbol{AB}$ produces only $O(1/p^2)$ memory instead of $O(p)$ as in other methods.



Intra-operator parallelism method: memory-efficient tensor parallelism

Method (Cont'd)

2. METP decomposes the computation of multi-head self-attention (MHA) into p^3 tasks and each device executes p^2 of them. The intermediate result $\boldsymbol{QK}^{\mathrm{T}}$ produces only $O(1/p^3)$ memory instead of $O(p)$ as in other methods.

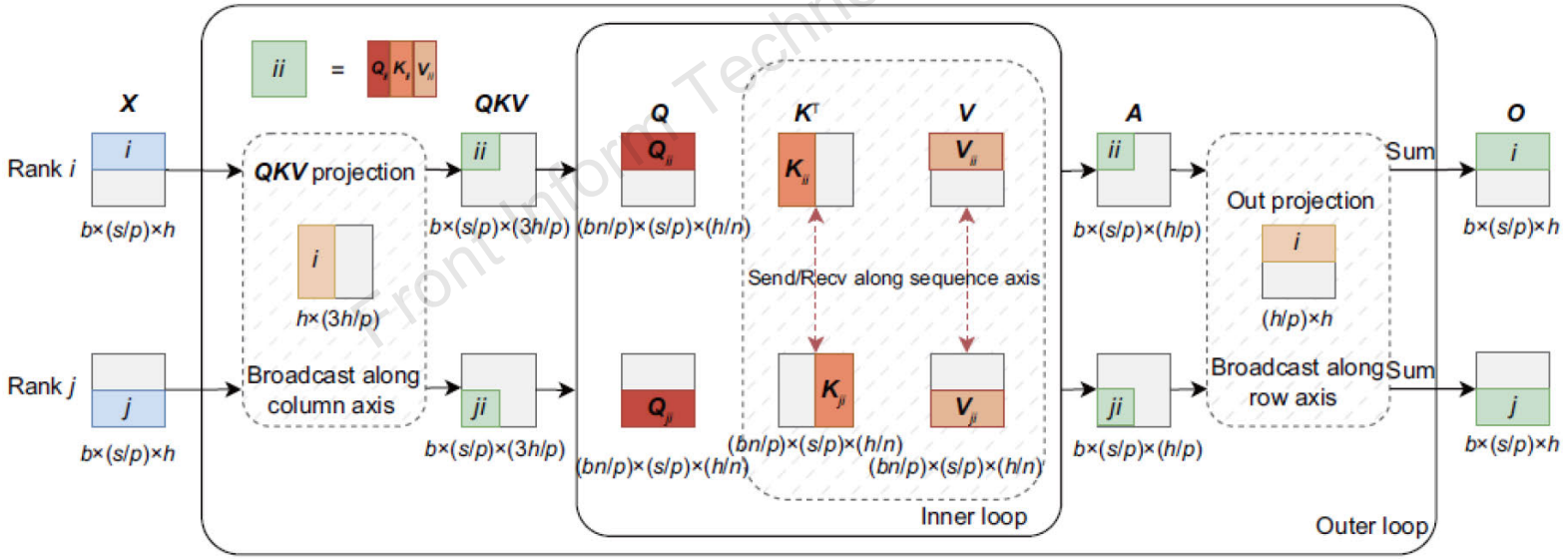


Fig. 2 METP for MHA (b: batch size; s: sequence length; h: hidden size; p: parallel degree; n: number of heads. Note: h/n=head_size. QKV will be transposed, reshaped, and split to obtain Q, K, V before they are fed into the inner loop. The workflow of the inner loop is similar to Fig. 1e)

Method (Cont'd)

3. To accelerate METP, we overlap the computation with communication of different iterations.

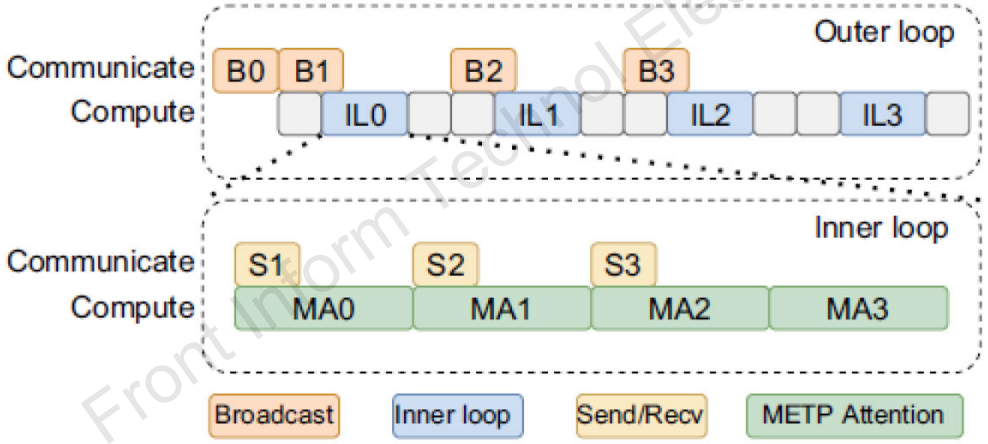


Fig. 3 Overlap between communication and computation in METP MHA

Major results

Compared with other methods that partition inputs along the sequence axis, METP achieves better memory efficiency.

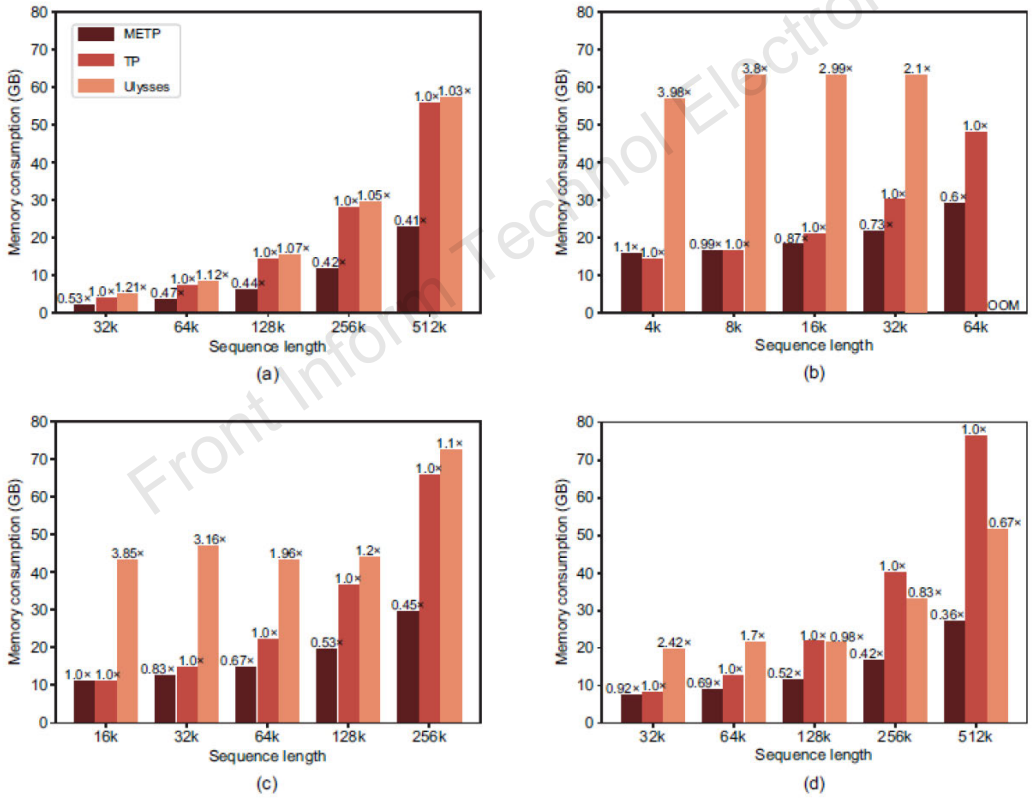


Fig. 4 Memory usage of different methods at different sequence lengths on eight A100 GPUs: (a) BERT-Large; (b) LLaMA-7B; (c) LLaMA-70B; (d) GPT-175B. Above each bar, we give their ratios compared to TP. OOM: out-of-memory

Major results (Cont'd)

METP can train longer sequences, while maintaining sufficient model FLOPS utilization (MFU).

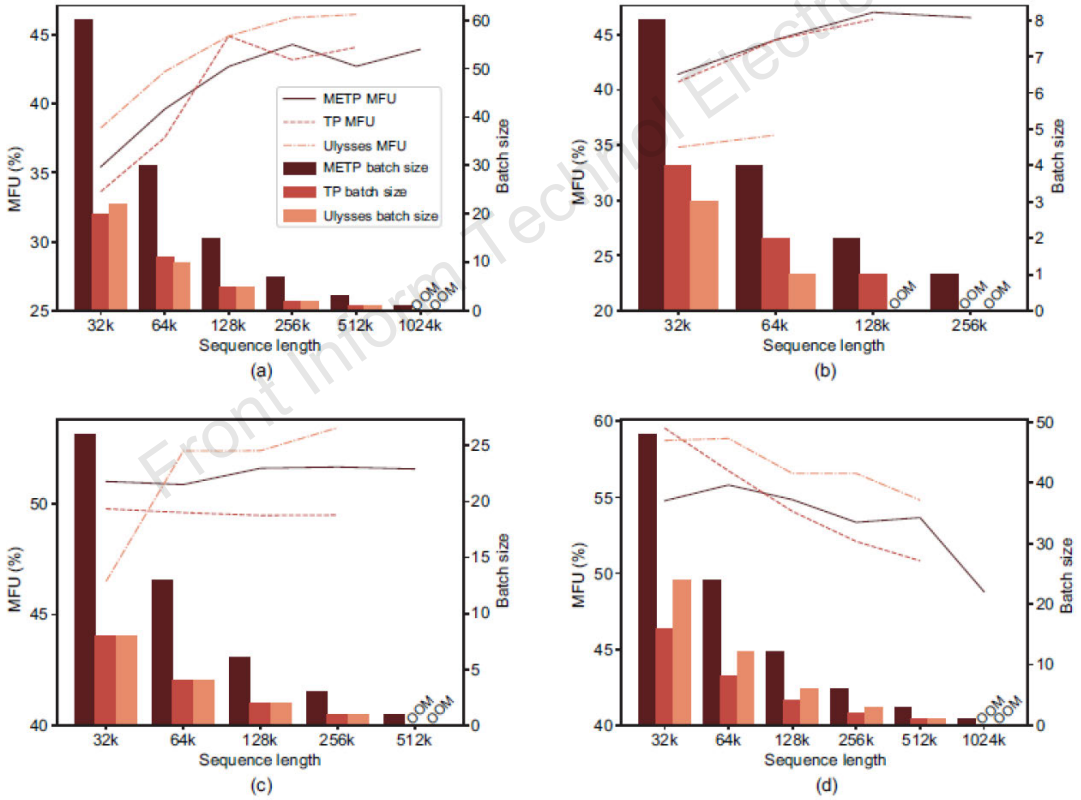


Fig. 5 MFU and the corresponding batch size for different models: (a) BERT-Large; (b) LLaMA-7B; (c) LLaMA-70B; (d) GPT-175B. OOM: out-of-memory

Major results (Cont'd)

METP has superlinear scalability on sequence length due to its memory savings.

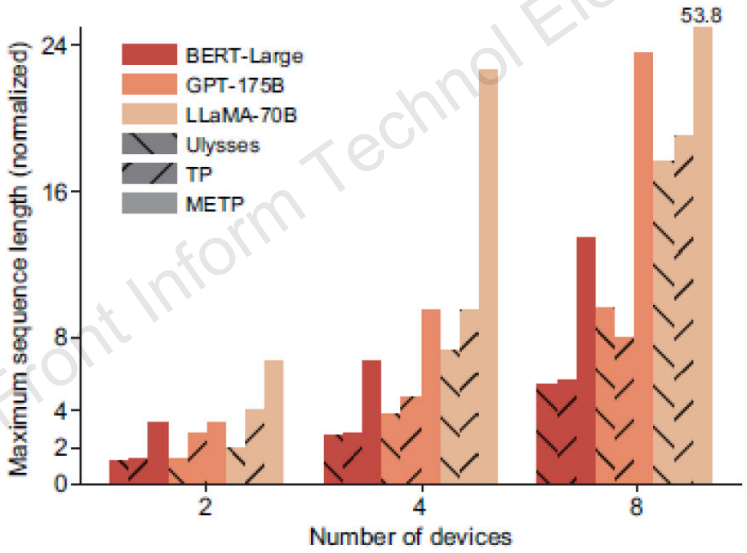


Fig. 6 Scaling sequence length by increasing the number of devices. Results are normalized by the maximum sequence length that a single device can achieve for each model setting

Conclusions

1. We propose a method named METP to optimize memory consumption during training LLMs with long sequences.

2. Our theoretical analysis demonstrates that METP can significantly reduce the memory consumption of attention to an $O(1/p^3)$ degree and reduce memory overheads of other intermediate results by at least 41.7%.

3. Our experimental results demonstrate that METP can increase the sequence length by 2.38–2.99 times compared to other methods.

Peng LIANG received the BS degree in network engineering from National University of Defense Technology (NUDT), Changsha, China, in 2019. He is currently working toward the PhD degree in computer science and technology at NUDT. His current research interests include deep learning and machine learning system.

Dongsheng LI received his PhD degree in computer science and technology from NUDT in 2005. He is currently a professor and doctoral supervisor in the College of Computer Science and Technology at NUDT. He was awarded the Chinese National Excellent Doctoral Dissertation in 2008. His research interests include distributed systems, cloud computing, and big data processing.