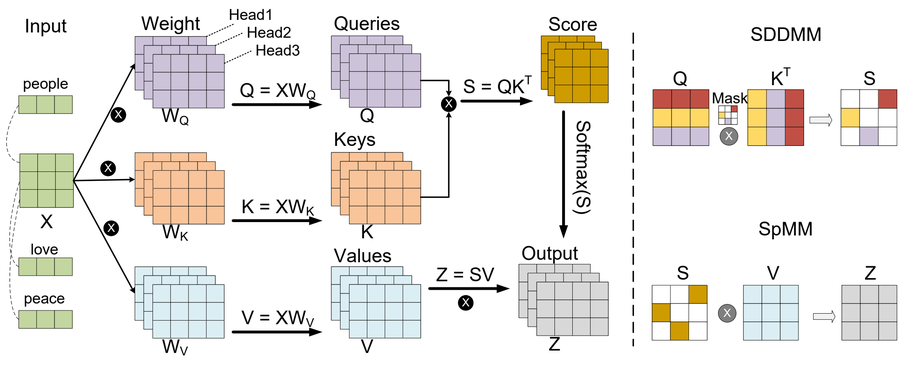
ASADI: Accelerating Sparse Attention using Diagonal-based In-situ Computing

Huize Li, **Zhaoying Li**, Zhenyu Bai, and Tulika Mitra

School of Computing, National University of Singapore

Basics of Sparse Attention

Transformer-based large models achieve state-of-the-art performance on various natural language processing and computer vision tasks. Researchers [2, 3] identity many connections in self-attention as weak connections, which can be eliminated to increase the execution efficiency with a slight loss of accuracy, namely sparse attention.



(a) Computation stages of dense attention mechanism (b) Sparse attention calculation

Fig. 1: Multi-head attention

Sparse Compression Format

The most common compression methods currently used are compress sparse row (CSR), compress sparse column (CSC), coordinate format (COO) and diagonal compression (DIA) format.

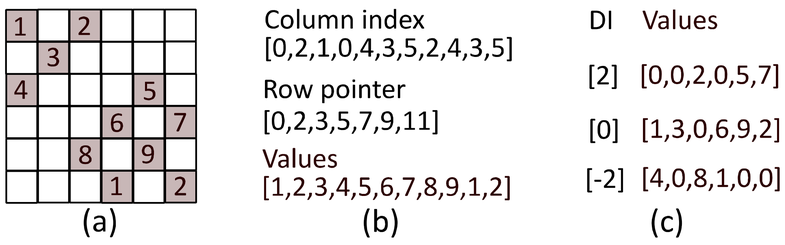


Fig. 2: (a) An example of sparse mask matrix, (b) CSR format, (c) DIA format

In-situ Computing

This paper focuses solely on ReRAM. ReRAM has the ability to perform two types of in-situ calculations: analog in-situ computing and digital in-situ computing.

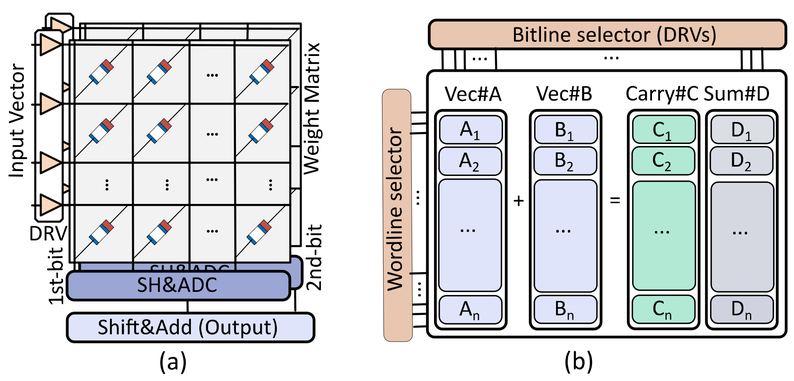


Fig. 3: (a) analog in-situ computing, and (b) digital in-situ computing

Motivations

- Observation#1: Diagonal locality is prevalent in static and dynamic sparse attention.
- Observation#2: PIM-based accelerators have high on-chip communication overhead.
- Our goal: Observation#1 motivates us to design a new matrix multiplication computation paradigm to efficiently support the DIA format. Observation#2 motivates us to design ASADI, minimizing on-chip transfers using in-situ computing.

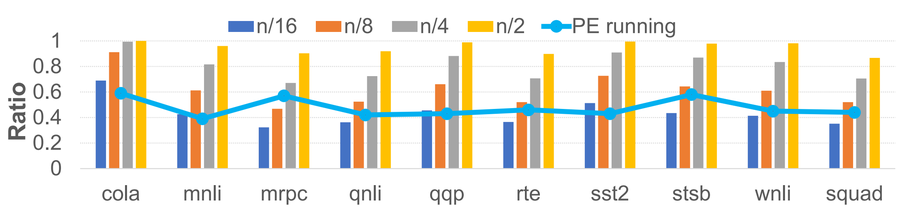


Fig. 4: Our observations

DIA-based In-situ Computing

In-situ $Q \times K^T$. Figure 5 (a) illustrates the sparse S matrix and dense V matrix. Figure 5 (b) presents the in-situ computation paradigm of the CSR format, where we assume that the CSR format of matrix S and the dense matrix V are stored in the same ReRAM array. The CSR storage format breaks the column coordinates of the left S matrix, preventing its direct use for in-situ matrix multiplication. Consequently, a row-wise remapping phase is necessary to align the coordinates.

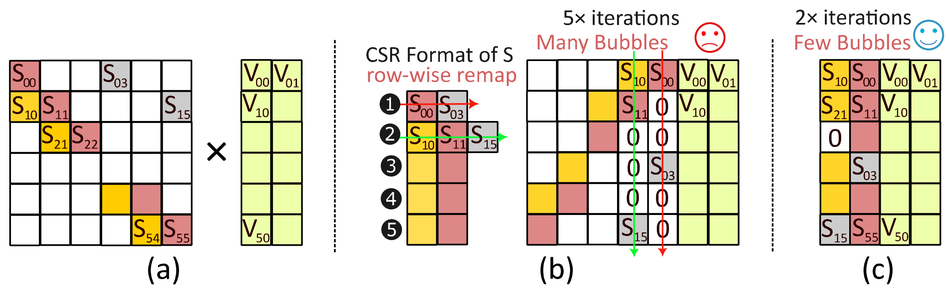


Fig. 5: SpMM computation paradigm

In-situ $S \times V$. Figure 6 (a) illustrates the SDDMM between dense matrices Q and K. Figure 6 (b) presents the SDDMM computation paradigm with the CSR format of matrix M. The ❶ iteration involves the first row of matrix M, controlling the 0-th row of matrix Q to calculate with the 0-th and 3-rd rows of matrix K. Since the ❶ and ❷ iterations share the 0-th row of matrix K, they must be executed serially. The CSR format takes five iteration with only two valid computing in each iteration.

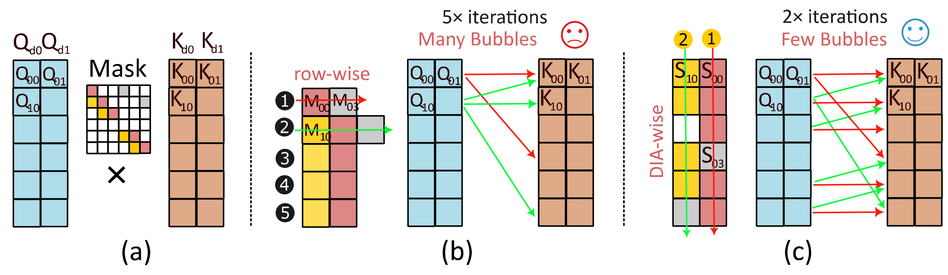


Fig. 6: SDDMM computation paradigm

Evaluation Results

Comparison with PIM baseline. ViL: 6.4× speedup; BERT: 2.3× to 63.7× on GLUE, SQuAD, WikeText, IMDB, Syn-4K, and Syn-8k datasets. BART: 1.9× to 60.1× speedups; GPT2: 2.1× to 61.7×.
Reasons: Reducing on-chip random access; The PIM baseline uses near-memory computation, where the on-chip logic units have many cross-bank data access.

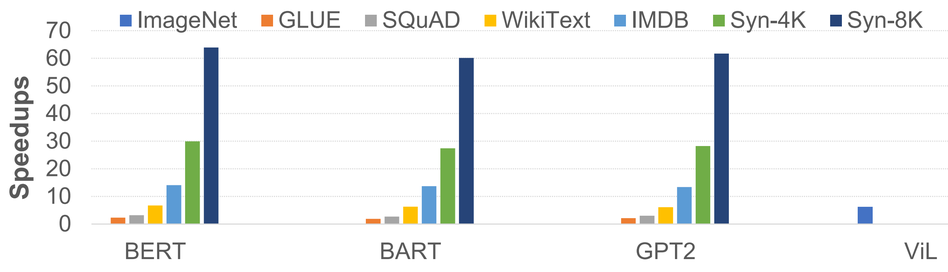


Fig. 7: Speedups compared to PIM baseline

Comparison with other platforms. GPU: RTX A6000; SPRINT: PIM for pruning and ASIC for attention [1]; CPSAA: PIM-based sparse attention accelerator.
Results and reasons: First, ASADI operates as a PIM-based accelerator, negating the need for extensive off-chip DRAM access. Second, ASADI supports the DIA format, which exhibits superior data locality, substantially mitigating random access demands. Finally, ASADI leverages in-situ computing hardware, featuring linear complexity growth with increasing sequence length.

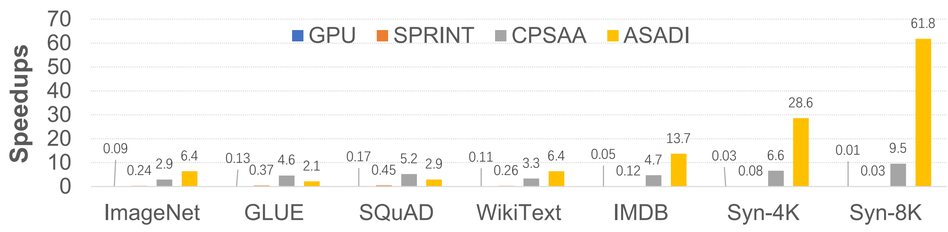


Fig. 8: Speedups compared to other platforms

References

[1] Amir Yazdanbakhsh et al. “Sparse Attention Acceleration with Synergistic In-Memory Pruning and On-Chip Recomputation”. In: Proceedings of 2022 IEEE/ACM International Symposium on Microarchitecture (MICRO). 2022, pp. 744–762.

[2] Iz Beltagy et al. “Longformer: The Long-Document Transformer”. In: CoRR abs/2004.05150 (2020). arXiv: 2004.05150. URL: https://arxiv.org/abs/2004.05150.

[3] Liqiang Lu et al. “Sanger: A Co-Design Framework for Enabling Sparse Attention Using Reconfigurable Architecture”. In: Proceedings of 54th Annual IEEE/ACM International Symposium on Microarchitecture. 2021, pp. 977–991. ISBN: 9781450385572.