Paper Review

Title: MulTCIM: Digital Computing-in-Memory-Based Multimodal Transformer Accelerator With Attention-Token-Bit Hybrid Sparsity

Conference: IEEE Journal of Solid-State Circuits, January 2024

Jiao Hongbao

Summary

This paper presents a digital computing-in-memory (CIM) accelerator tailored for multimodal Transformers, focusing on exploiting data sparsity to improve energy efficiency. The main contributions include:

1. **LRES (Long Reuse Elimination Scheduler)**: Reshapes the attention matrix into global + local patterns to retain frequently accessed data in the CIM buffer, improving CIM utilization in StageD.
2. **RTP (Runtime Token Pruner) & MACN (Modal-Adaptive CIM Network)**: Prunes unimportant tokens in advance and leverages MACN to divide CIM into StageS and StageD. It performs pipeline scheduling by exploiting symmetrical computation overlap between modalities, reducing CIM idleness.
3. **EBB-CIM (Effective-Bit Balanced CIM)**: Performs bit-level analysis and balances bit usage to mitigate resource waste due to uneven effective bit widths, thereby reducing computation latency.

Strengths

- The three architectural designs—LRES, RTP+MACN, and EBB-CIM—effectively target real-world circuit bottlenecks. The system captures sparsity at attention, token, and bit levels throughout the QKV computation of Transformer models.
- The paper includes a comprehensive pipeline from problem identification to hardware validation, with convincing experimental results based on a fabricated chip.
- The topic is timely and aligned with current trends in energy-efficient AI accelerators for large-scale multimodal models.

Weaknesses

- The experiments are primarily based on the ViLBERT model. The applicability of the architecture to larger-scale multimodal models

(e.g., Flamingo) is not thoroughly discussed, which affects the generality of the results.

- The paper lacks quantitative analysis of the pre-processing cost associated with handling sparsity. The energy or latency overhead for sparsity detection is not explicitly addressed.

Suggestions for Improvement

- Expand the experimental scope to include more modern and diverse multimodal models with varying sparsity patterns.
- Provide additional analysis on the overhead introduced by sparsity detection and handling mechanisms, especially energy and latency impact.
- Since all sparsity handling is done at the hardware level, consider exploring compiler-hardware interface design to allow software-level sparsity scheduling and metadata injection.

Overall Evaluation

This paper proposes a Transformer accelerator that exploits sparsity at multiple levels (attention, token, bit) with architectural innovations that are both practical and novel. It provides sufficient experimental validation and holds strong publication value.

论文评审

Title: MulTCIM: Digital Computing-in-Memory-Based Multimodal Transformer Accelerator With Attention-Token-Bit Hybrid Sparsity

Conference: IEEE Journal of Solid-State Circuits January 2024

简要总结

本论文提出了一种为多模态Transformer设计的数字存内计算加速器，主要实现了针对数据的稀疏性优化，核心贡献包括以下三个方面:

1. **LRES (Long Reuse Elimination scheduler)**: 将注意力矩阵重构为全局+局部 的结构，使得被频繁使用的数据保存到CIM中，提高StageD的CIM使用率。
2. **RIP (Runtime Token Pruner) && MACN (Modal-Adaptive CIM Network)**: 提前去除无关紧要的token，并通过MACN将CIM划分为StageS和StageD实现模态负载调度，通过模态之间计算的对称重叠设计流水线降低CIM的空闲率。

3. **EBB-CIM (effective-Bit Balanced CIM)**：通过有效位检测和bit平衡防止不同位宽和有效位差距过大造成的资源浪费，并降低计算时间。

优点

- LRES,RTP+MACN,EBB-CIM，三个电路架构设计针对当下电路实际瓶颈问题，目标明确，针对数据的稀疏性，能够从Transformer模型的QKV计算全流程，从attention，token,bit 三个层级进行加速设计。
- 论文包含从目标问题，设计思路，到芯片测试的全流程，包含实际数据，说服力强。
- 选题符合趋势，面向大规模多模态模型的能效优化，符合当下AI电路发展热点。

不足之处

- 目前模型主要面向ViBERT模型，对于当前更大的多模态模型（如Flamingo等）的适配性探讨不足，评估基准较为了结果的推广性。
- 论文对于数据稀疏性的前期处理开销的量化分析不足，未能明确提出这方面的额外开销分析。
-

改进建议

- 可以扩展实验模型范围，用更多更新的模型（不同模态和稀疏性）进行测试分析。
- 补充对于稀疏性处理机制带来的能耗与开销成本分析。
- 稀疏性判断及其处理由硬件完成，可以考虑建立编译器和实际电路的逻辑接口。

总评

本论文是关于多模态transformer加速器的相关研究，在Attention-Token-Bit的多个层面对数据稀疏性进行了讨论分析和相应处理和计算逻辑设计，具备创新价值和实验证据，具备发表的价值。