先端科学技術研究科　修士論文要旨

所属研究室 (主指導教員)	コンピューティング・アーキテクチャ (中島　康彦 (教授))		
学籍番号	2411169	提出日	令和 8年 1月 16日
学生氏名	竹内　歩夢		
論文題目	Implementation and Evaluation of Quantization-Free Attention Computation on a CGLA Accelerator CGLAアクセラレータにおける非量子化Attention計算の実装と評価		

要旨

With the rapid expansion of generative AI, Transformer models have become foundational to modern AI systems. However, attention remains a dominant bottleneck because both compute and memory footprint grow quadratically with sequence length. FlashAttention addresses this challenge by restructuring attention into block-wise computation that reduces off-chip memory traffic while remaining mathematically equivalent to standard attention. Despite its impact, FlashAttention was developed for NVIDIA GPUs, and practical, highly optimized implementations have been largely confined to GPU platforms.

This thesis advances power-efficient attention beyond GPUs by realizing FlashAttention on IMAX3, a Coarse-Grained Linear Array (CGLA) accelerator. We redesign FlashAttention's blocking strategy and memory layout to match IMAX3's resource constraints and data-movement characteristics, enabling efficient attention execution on a linear-array accelerator. Experimental results demonstrate a 4-5× speedup over naive attention, reduced memory usage, and preserved numerical accuracy. Furthermore, an Energy-Delay Product (EDP) study shows that, for token lengths of 500 or less, FlashAttention on IMAX3 is more power-efficient than an NVIDIA RTX 4090, achieving up to a 472× EDP improvement.

Beyond the port of a GPU-centric algorithm, this thesis also introduces an IMAX3-native attention approach for LLM inference that minimizes reliance on the host ARM processor. In IMAX3-based deployments, non-trivial overhead can arise when operations that cannot be executed on IMAX3 fall back to the host ARM processor. We propose an attention formulation and execution strategy designed to avoid such ARM-side work as much as possible.