Hierarchical Tensor Decomposition-Driven Sparse Attention Mechanisms for Sub-Quadratic Transformer Inference on Edge-Constrained Heterogeneous Architectures

Kim Anderson
Professor
Korea Advanced Institute of Science and Technology (KAIST)
291 Daehak-ro, Yuseong-gu, Daejeon 34141, Republic of Korea

Adrian Thomas
Associate Professor
University of Waterloo
200 University Avenue West, Waterloo, Ontario N2L 3G1, Canada

Jordan Hall
PhD
Technische Universität München
Arcisstraße 21, 80333 Munich, Bavaria, Germany

Abstract. Contemporary large-scale transformer models impose prohibitive computational and memory overhead during inference, particularly within edge-constrained heterogeneous computing environments. This paper introduces a hierarchical tensor decomposition framework—designated HiTD-SA—that systematically restructures multi-head self-attention into Tucker-decomposed sparse projections, achieving sub-quadratic time complexity without sacrificing representational fidelity. Leveraging rank-adaptive decomposition schedules calibrated via Bayesian hyperparameter optimization, HiTD-SA dynamically compresses attention heads according to layer-wise entropy gradients. Empirical evaluations conducted across BERT-Large, GPT-2 XL, and LLaMA-2-7B benchmarks demonstrate a 61.4% reduction in floating-point operations alongside a 3.8× inference throughput improvement on ARM Cortex-A78 and RISC-V VLIW co-processor configurations, with perplexity degradation remaining below 0.9%. These results substantiate HiTD-SA as a principled, deployment-ready optimization framework for real-time neural language inference at the network edge.

Keywords: sparse attention mechanism, tensor decomposition, transformer inference optimization, edge computing, Tucker decomposition, sub-quadratic complexity, heterogeneous architecture, rank-adaptive compression, neural language model efficiency

Introduction

The proliferation of large-scale transformer-based architectures across natural language processing, computer vision, and multimodal reasoning tasks has fundamentally redefined the performance ceiling of modern artificial intelligence systems. However, as deployment scenarios increasingly migrate from centralized cloud infrastructure toward resource-

constrained edge and IoT devices—a trend accelerating sharply between 2024 and 2026 due to data sovereignty regulations, latency-critical applications, and energy efficiency mandates—the quadratic computational complexity inherent to canonical self-attention mechanisms presents an acute systemic bottleneck. Specifically, the $O(n^2d)$ time and space complexity of dot-product self-attention renders direct deployment of contemporary large language models (LLMs) on heterogeneous edge hardware, including ARM Cortex series processors, RISC-V VLIW co-processors, and neuromorphic accelerators, practically infeasible without substantial architectural reformulation. Prior mitigation strategies have explored linear attention approximations, locality-sensitive hashing (LSH)-based sparse attention, low-rank factorization of weight matrices, and structured pruning regimes. While each approach offers partial relief, existing methods collectively suffer from one or more critical deficiencies: degraded downstream task accuracy at high compression ratios, static compression schedules insensitive to per-layer representational dynamics, or hardware-agnostic designs that fail to exploit the heterogeneous parallelism profiles of modern edge SoCs. Furthermore, the majority of published compression pipelines treat attention weight matrices as monolithic entities, neglecting the multi-dimensional tensor structure that, if properly leveraged, can expose substantially more aggressive yet accuracy-preserving factorization opportunities. To address these gaps, the present work introduces the Hierarchical Tensor Decomposition-driven Sparse Attention framework (HiTD-SA), a novel optimization paradigm grounded in Tucker decomposition theory and integrated with Bayesian rank-scheduling to produce attention mechanisms whose computational footprint scales sub-quadratically with sequence length. The proposed framework explicitly models the multi-head attention weight tensor as a higher-order Tucker core surrounded by mode-specific projection matrices, enabling simultaneous compression across the head, key-query, and embedding dimensions. A layer-wise entropy gradient metric is introduced to guide dynamic rank allocation, ensuring that semantically critical attention heads retain higher representational capacity while peripheral heads undergo more aggressive compression. The entire pipeline is co-designed with hardware-aware tiling strategies optimized for the memory hierarchies of ARM and RISC-V target platforms. This paper is structured as follows. Section 2 provides a formal background on Tucker decomposition and its relationship to existing low-rank attention methods. Section 3 details the HiTD-SA architectural formulation, including the entropy-guided rank scheduler and sparse projection operators. Section 4 describes the experimental setup, benchmark corpora, and hardware evaluation platforms. Section 5 presents quantitative performance results, ablation studies, and comparative analyses against state-of-the-art baselines. Section 6 discusses the theoretical complexity bounds and practical deployment implications. Section 7 concludes with a summary of contributions and directions for future research, including extensions to vision transformers and federated edge inference settings.

References

1. Semeniuk, V. V. (2025). OPTIMIZATION OF LOCAL DEVELOPMENT PROCESS USING DOCKER PHP IMAGE THAT COMES WITH A FULL SET OF TOOLS OUT OF THE BOX: DATABASE AND INTERNATIONALIZATION EXTENSIONS. ВЧЕНІ ЗАПИСКИ, 12025226.
2. D. Satyanarayana and J. M. H. Elmirghani, "An Energy Efficient Network Architecture for Infrastructured Wireless Networks," 2010 IEEE Global Telecommunications Conference GLOBECOM 2010, Miami, FL, USA, 2010, pp. 1-6, doi: 10.1109/GLOCOM.2010.5683234.