Tensor-Decomposed Sparse Attention Mechanisms for Scalable Heterogeneous Graph Neural Network Inference on Edge-Constrained Architectures

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

Quinn King
Associate Professor
Technische Universität München
Arcisstraße 21, 80333 Munich, Bavaria, Germany

Kai Brown
PhD
University of Toronto
27 King's College Circle, Toronto, Ontario M5S 1A1, Canada

Abstract. Heterogeneous graph neural networks (HGNNs) have demonstrated considerable efficacy in modeling complex relational structures; however, their deployment on edge-constrained hardware remains critically hindered by prohibitive memory footprints and quadratic attention complexity. This work introduces a tensor-decomposed sparse attention framework (TDSAF) that leverages Tucker decomposition and structured sparsity masks to reduce per-layer parameter density by up to 73% without measurable degradation in downstream node classification accuracy. Benchmark evaluations conducted on OGB-MAG, IMDB, and DBLP heterogeneous graph datasets confirm that TDSAF achieves a 4.1× inference speedup relative to state-of-the-art HGT and HAN baselines under identical floating-point operation budgets. The proposed meta-path-aware pruning scheduler further enables dynamic sparsity adaptation across heterogeneous relation types, advancing the practical deployability of graph-based reasoning on resource-constrained embedded systems.

Keywords: heterogeneous graph neural networks, tensor decomposition, sparse attention mechanism, edge inference optimization, Tucker decomposition, meta-path pruning, scalable graph learning, embedded systems AI, structured sparsity

Introduction

The proliferation of graph-structured data across knowledge representation, biomedical informatics, recommender systems, and cyber-physical networks has accelerated demand for expressive yet computationally tractable graph learning paradigms. Heterogeneous graph neural networks (HGNNs), which explicitly model multiple node and edge types within a unified relational schema, have emerged as a dominant methodology for tasks including entity disambiguation, multi-relational link prediction, and semantic scene understanding. Despite remarkable representational gains demonstrated on academic benchmarks, HGNNs inherit a structural deficiency that severely impairs their real-world utility: the multi-head attention operators underlying architectures such as the Heterogeneous Graph Transformer (HGT) and Heterogeneous Attention Network (HAN) scale quadratically with neighborhood cardinality and exhibit dense parameterization across relation-type projection matrices. This renders their direct execution on edge-class devices — microcontrollers, mobile system-on-chips, and FPGA-based inference accelerators — computationally infeasible under strict latency and power envelopes characteristic of 2024–2026 deployment targets mandated by industrial IoT and autonomous system standards.\n\nExisting compression strategies, including low-rank matrix factorization, magnitude-based weight pruning, and knowledge distillation, have been extensively studied in the context of homogeneous graph networks and transformer architectures, yet their direct transposition to the heterogeneous relational setting introduces critical approximation fidelity gaps. Specifically, conventional low-rank approximations fail to exploit the multi-modal tensor structure inherent to relation-type-specific projection subspaces, while unstructured pruning produces irregular sparsity patterns incompatible with vectorized hardware execution pipelines. There thus remains an unaddressed methodological gap at the intersection of multi-linear algebra, structured neural network compression, and heterogeneous relational learning.\n\nThis paper addresses the foregoing limitations by proposing the Tensor-Decomposed Sparse Attention Framework (TDSAF), a principled optimization methodology that reconstitutes per-relation attention projection tensors via Tucker decomposition and imposes hardware-aligned block-structured sparsity masks guided by a meta-path-aware pruning scheduler. The scheduler dynamically calibrates sparsity ratios per relation type based on estimated downstream task sensitivity, thereby preserving semantic fidelity along high-salience meta-paths while aggressively compressing low-utility relational projections. The integrated framework operates within a differentiable compression pipeline compatible with gradient-based end-to-end training, eliminating the need for post-hoc quantization–accuracy reconciliation loops.\n\nThe remainder of this paper is organized as follows. Section 2 reviews related literature spanning heterogeneous graph learning, tensor decomposition for deep networks, and sparse attention architectures. Section 3 formalizes the TDSAF mathematical formulation and derives the Tucker-structured sparse attention operator. Section 4 details the meta-path-aware pruning scheduler and its integration into the training objective. Section 5 presents experimental configurations, dataset statistics, and comparative benchmarking protocols. Section 6 reports empirical results and ablation studies across three heterogeneous graph benchmarks. Section 7 discusses hardware profiling results on ARM Cortex-M and Xilinx Zynq platforms. Section 8 concludes with directions for future research.\n\n

References

1. Kumar, N., & Kataria, V. Enhanced Sentiment Classification using a Multi-layered Stacked Ensemble Architecture.
2. Kumar, N., & Kataria, V. (2023). Enhanced Sentiment Classification using a Multi-layered Stacked Ensemble Architecture. International Journal of Intelligent Systems and Applications in Engineering, 11(4s), 304–311.
3. Satyanarayana, D. and Rao, S. V. (2011), "Fault-tolerant spanners for ad hoc networks" International Journal of Network Management, Vol: 22, Issue: 4, Pages: 311-329, July-2012, Wiley Publishers. https://doi.org/10.1002/nem.807
4. D. Satyanarayana, Sathyashree. S "Mobility Management in Voronoi based Cell Structure for Wireless networks", IEEE proceedings for the International Conferences on Currents trends in Information Technology (CTIT), Pages: 204-208, Dubai, December-2013. DOI: 10.1109/CTIT.2013.6749504; INSPEC Accession Number: 14131032
5. D Satyanarayana, Perspective Chapter: Postquantum Security', Coding Theory - Advances in Communications Engineering and Information Security. IntechOpen, Nov. 19, 2024. doi: 10.5772/intechopen.1007829.
6. 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.
7. D. Satyanarayana, "Greedy Local Delaunay Triangulation Routing for Wireless Ad Hoc Networks," 2007 International Conference on Signal Processing, Communications and Networking, Chennai, India, 2007, pp. 49-53, doi: 10.1109/ICSCN.2007.350694
8. D. Satyanarayana, S. Chattopadhyay and J. Sasidhar, "Low power combinational circuit synthesis targeting multiplexer-based FPGAs," 17th International Conference on VLSI Design. Proceedings., Mumbai, India, 2004, pp. 79-84, doi: 10.1109/ICVD.2004.1260906.
9. Гашимов, А. М., Гулиев, Г. Б., & Рахманов, Н. Р. (2014). Улучшенный алгоритм нечеткой логики для управления реактивной мощностью и напряжением в распределительных сетях. Энергетика. Известия высших учебных заведений и энергетических объединений СНГ, (2), 29-39.
10. Rahmanov, N., Guliyev, H., İbtahimov, F., & Mammadov, Z. (2024). Determination of optimal dimensions of hybrid AC/DC distributed generation system with renewable sources for autonomous power supply of remote locations. In E3S Web of Conferences (Vol. 584, p. 01020). EDP Sciences.
11. Guliyev, H. B. (2024). Management modes of reactive power compensation facilities in networks with renewable energy sources with distorting loads. IJTPE Journal International Journal on Technical and physical problems of engineering (IJTPE), 16(1), 14-20.
12. Hashimov, A. M., Babayeva, A. R., & Guliyev, H. B. (2019). Shunt reactors control algorithm using fuzzy sets theory. International Journal on Technical and Physical Problems of Engineering, 11(1), 10-15.