ELEVENTH PRESS INTERNATIONAL PUBLISHING™
Eleventh Press International Journal of Multidisciplinary Research
ISSN 2945-1138 · Vol. 4, No. 2 (2024) · DOI: 10.52011/epip.2024.8096

RESEARCH ARTICLE · COMPUTER SCIENCE

Sparse Mixture-of-Experts Routing for Resource-Constrained Edge Inference

Liu Chen[1], Priya Raghunathan[2], Sven Larsson[3]

[1] Tsinghua University
[2] Indian Institute of Science
[3] KTH Royal Institute of Technology

ABSTRACT

Mixture-of-experts (MoE) architectures have become a standard technique for scaling transformer models without proportional increases in inference cost, yet their deployment on commodity edge accelerators remains hampered by irregular memory access patterns and load imbalance. We introduce SparseEdge-MoE, a routing mechanism that combines locality-sensitive hashing with hierarchical expert grouping to reduce expert activation variance by 41% on a Raspberry Pi 5 platform. Benchmarked against GLaM and Mixtral configurations, SparseEdge-MoE attains within 2.1% of dense baseline accuracy on SuperGLUE while consuming 3.8x less peak memory bandwidth. We release the runtime and benchmark suite under a permissive open-source license.

Keywords: *mixture of experts, edge inference, transformers, model compression, sparse routing*

Cite as: Chen, L., Raghunathan, P., & Larsson, S. (2024). Sparse Mixture-of-Experts Routing for Resource-Constrained Edge Inference. Eleventh Press International Journal of Multidisciplinary Research, 4(2). https://doi.org/10.52011/epip.2024.8096

Licence: © 2024 the authors. Open access under the Creative Commons Attribution 4.0 International licence (CC BY 4.0), which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

Authors

- **Liu Chen** — Tsinghua University (ORCID 0000-0002-4444-0001)
- **Priya Raghunathan** — Indian Institute of Science
- **Sven Larsson** — KTH Royal Institute of Technology

Abstract

Mixture-of-experts (MoE) architectures have become a standard technique for scaling transformer models without proportional increases in inference cost, yet their deployment on commodity edge accelerators remains hampered by irregular memory access patterns and load imbalance. We introduce SparseEdge-MoE, a routing mechanism that combines locality-sensitive hashing with hierarchical expert grouping to reduce expert activation variance by 41% on a Raspberry Pi 5 platform. Benchmarked against GLaM and Mixtral configurations, SparseEdge-MoE attains within 2.1% of dense baseline accuracy on SuperGLUE while consuming 3.8x less peak memory bandwidth. We release the runtime and benchmark suite under a permissive open-source license.

1. Introduction

This article was published by Eleventh Press International Journal of Multidisciplinary Research under a single blind peer-review model. The full text of the manuscript has been deposited alongside this record and is available via the galley links.

2. Methods

Methodological detail is preserved verbatim from the accepted manuscript. The production service has rendered this HTML, PDF, and JATS galley using Pandoc 3.1.11.1 and WeasyPrint with the journal's house CSS template.

3. Results

The findings, figures, and tables of the original submission are reproduced here. Readers should consult the PDF galley for the authoritative typeset version.

4. Discussion

We introduce SparseEdge-MoE, a routing mechanism that combines locality-sensitive hashing with hierarchical expert grouping to reduce expert activation variance by 41% on a Raspberry Pi 5 platform

References

1. Mauduit, C. & Rivat, J. (2009). La somme des chiffres des nombres premiers. *Annals of Mathematics*, 171(3), 1591–1646.
2. Patel, M. et al. (2023). Strain engineering in 2D materials. *Nature Reviews Physics*, 5, 412–429.

Keywords: mixture of experts, edge inference, transformers, model compression, sparse routing Discipline: Computer Science