INNOVATION PARTNERSHIPS
UNIVERSITY OF MICHIGAN

GRACE: A Scalable Graph-Based Approach to Accelerating Recommendation Model Inference

TECHNOLOGY NUMBER: 2024-370

OVERVIEW

Graph-based algorithm for enhanced recommendation model performance in data centers

- Reduces memory traffic in deep learning models using graph co-occurrence analysis
- Enhances efficiency for streaming services, e-commerce, and advertising platforms

BACKGROUND

Deep learning recommendation models (DLRMs) are crucial for tailoring content such as news feeds and product suggestions across vast networks like those operated by Meta, Google, and Amazon. Traditionally, these models run on collaborative CPU-GPU systems to optimize computational capacities. Despite architectural advancements, current methods incur high memory bandwidth usage, particularly due to DLRMs' vast embedding layers. This leads to significant performance bottlenecks and increased data center operational costs. Historical techniques, such as partial sum caching or employing heterogeneous memory systems, offer limited scalability and benefits as they either handle restricted embedding subsets or exhibit high processing costs. Therefore, there's a compelling need for scalable solutions that reduce data traffic while maintaining high-speed model inference in data center environments.

INNOVATION

Researchers at the University of Michigan have developed, GRACE, which introduces a scalable graph-based framework that leverages item co-occurrence patterns to advance recommendation model inference. By constructing an Item Co-occurrence Graph (ICG), the system identifies and clusters frequently co-accessed items to pre-compute and store their partial sums. This approach minimizes the memory traffic seen with sparse embedding layer operations. The innovation lies in its system-aware design, integrating seamlessly into existing DRAM-based systems to enable significant throughput improvement without additional hardware requirements. Potential applications extend across data-centered operations, optimizing recommendation engines for streaming services, commercial retailers, and online advertisement platforms, ultimately facilitating rapid personalization and enhanced user engagement in real time.

ADDITIONAL INFORMATION

Technology ID
2024-370

Category
Hardware
Engineering & Physical Sciences
Semiconductor, MEMS, and Electronics

Inventor
Nishil Rakeshkumar Talati
Haojie Ye
Yuhan Chen
Yichen Yang
Trevor Mudge
Ronald Dreslinski Jr.
Alex Bronstein
Sanketh Vedula

Further information
Joohee Kim
jooheek@umich.edu

View online

REFERENCES

Ye, Haojie and Vedula, Sanketh and Chen, Yuhan and Yang, Yichen and Bronstein, Alex and Dreslinski, Ronald and Mudge, Trevor and Talati, Nishil, "GRACE: A Scalable Graph-Based Approach to Accelerating Recommendation Model Inference", Proceedings of the 28th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 3, 2023, doi:10.1145/3582016.3582029

INTELLECTUAL PROPERTY

Pending