MnnFast: A Fast and Scalable System Architecture for Memory-Augmented Neural Network

Corresponding author

Hanhwi Jang, Jae-Eon Jo, POSTECH Pohang, Republic of Korea

Joonsung Kim, Jaewon Lee, Jangwoo Kim, Seoul National University, Republic of Korea

Keywords

Memory Networks, Attention-based Neural Networks, Machine Learning, Parallel Algorithm, Computation/Dataflow Optimization, Accelerator, Algorithm-Hardware Co-Design, Architecture

Summary

*Challenge*

As the size of input datasets rapidly grows, the current computer infrastructure cannot achieve scalable performance due to its limited system architecture. They identify the performance problems of the current architecture by conducting extensive performance bottleneck analysis. It indicates that the current architecture suffers from three major performance problems: **high memory bandwidth consumption, heavy computation, and cache contention**.

*Contribution*

1. Propose **MnnFast**, a novel system architecture for large-scale memory networks to achieve fast and scalable reasoning performance.

1). Propose a new **column-based algorithm** with streaming to reduce the memory bandwidth consumption.

2). Propose a **zero-skipping optimization** to bypass a large amount of output computation which significantly decrease the high computational overhead.

3). Propose an **embedding cache** dedicated to efficiently cache the embedding matrix which designed to eliminate the cache contention.

1. Implement the baseline MemNN and MnnFast on CPU, GPU and FPGA.

*Innovation Points*

1. The column-based algorithm partitions input/output memory into multiple chunks, and calculates partial output vectors for each chunk.

*Result*

MnnFast is significantly effective in various types of hardware: CPU, GPU, and FPGA. MnnFast improves the overall throughput by up to 5.38×, 4.34×, and 2.01× on CPU, GPU, and FPGA respectively. Also, compared to CPU-based MnnFast, our FPGA-based MnnFast achieves 6.54× higher energy efficiency.