Acceleration of SpMM for GNN Inference with CGLA

Name Koki Asahina

Laboratory Computing Architecture

Supervisor Prof. Yasuhiko Nakashima

Abstract

Sparse matrix–matrix multiplication (SpMM) is a core operation in Graph Neural Networks (GNNs) but often becomes the main performance bottleneck. In edge computing environments with limited compute and memory resources, highly sparse adjacency matrices suffer from poor cache locality, degrading both performance and energy efficiency. This calls for hardware-level, dataflow-oriented optimization.

We propose IMAX-SpMM, a method that accelerates SpMM while improving energy efficiency. It runs on In-Memory Accelerator eXtension 3 (IMAX3), which adopts a Coarse-Grained Linear Array (CGLA) architecture to achieve high data reuse and efficient computation. Node feature vectors are grouped by their number of nonzero elements, and each group is processed with kernels whose lengths are tuned to the vector size. In addition, instructions are remapped to processing elements (PEs) on a per-dataset basis. Together, these techniques yield a 60.1% data reuse rate for Direct Memory Access (DMA) transfers, greatly improving memory transfer efficiency.

On Graph Convolutional Networks (GCNs), the proposed method achieves up to 9.15× speedup over an i9-10940X CPU and 4.89× over an RTX3090 GPU. In terms of energy efficiency, it improves performance per watt by up to 3,089× compared to the i9-10940X and 321× compared to the RTX3090. The performance and power figures for the ASIC implementation are estimated from frequency- and bandwidth-scaled FPGA measurements and synthesis results using a 28 nm standard cell library, and therefore are not guaranteed for actual silicon.