先端科学技術研究科　修士論文要旨

所属研究室 (主指導教員)	コンピューティング・アーキテクチャ (中島　康彦　(教授))		
学籍番号	2311006	提出日	令和 7年 1月 17日
学生氏名	朝比奈　甲樹		
論文題目	Energy-Efficient SpMM kernels for GCNs and GATs on a CGLA		

要旨

Graph Neural Networks (GNNs) have demonstrated excellent performance in various fields, such as atomic structure estimation, chemical reaction modeling, and information retrieval support in Large Language Models (LLMs). In particular, Graph Convolutional Networks (GCNs) and Graph Attention Networks (GATs) have been widely applied due to their computational efficiency and scalability. However, Sparse Matrix-Matrix Multiplication (SpMM) is a bottleneck for these models. CPUs, GPUs, and hardware accelerators struggle to balance high performance and low power consumption, making them unsuitable for energy-efficient graph applications. To solve this problem, we propose SpMM kernels that utilize a Coarse-Grained Reconfigurable Array (CGRA)-based accelerator, IMAX3 (In-Memory Accelerator eXtension 3), which is a power-efficient reconfigurable accelerator. This study introduces a new data format focused on minimizing the number of Direct Memory Access (DMA) communications. In addition, it proposes an optimized instruction mapping for seamless operations on IMAX3. When implemented on a Xilinx VPK180 FPGA, IMAX3 achieves up to 6.44x speedup over the i9-10940X, 10.04x over the RTX3090, and 46.26x over the Jetson AGX Orin. Furthermore, when estimated for ASIC implementation using TSMC's 28nm process library, IMAX3 demonstrates Energy-Delay Product (EDP) improvements ranging from 20.24x to 1,334.12x over the i9-10940X, up to 1.54x over the RTX3090, and up to 172.31x over the Jetson AGX Orin.