Memory Transfer-Aware Acceleration of Similarity Search on CGLAs

Name Dohyun Kim

Laboratory Computing Architecture

Supervisor Prof. Yasuhiko Nakashima

Abstract (should be within 1st page)

Similarity search is a fundamental technique for retrieving information from unstructured data and has been widely applied in domains such as image, speech, and text retrieval. Its significance has further increased with the rise of Retrieval-Augmented Generation (RAG) in Large Language Models (LLMs). However, similarity search is computationally intensive, and even with algorithmic advances in approximate methods, the demand for high computational and memory resources remains.

This study aims to accelerate similarity search using IMAX3, a type of Coarse-Grained Reconfigurable Linear Array (CGLA). IMAX3 features large local memory within each processing unit, enabling efficient reuse of data and reduced data movement. We focus on optimizing k-nearest neighbor (k-NN) search by exploiting IMAX3's architectural characteristics to improve memory access patterns. The proposed method achieves up to 54.9× speedup for repeated queries and up to 498.41× reduction in Energy-Delay Product (EDP) compared to RTX 4090, demonstrating its suitability for real-time vector search applications.

Furthermore, we apply similar optimization strategies to Graph Convolutional Networks (GCNs), which process sparse graph-structured data. By reducing redundant memory operations and optimizing access patterns, our implementation achieves an average of 3.64× higher energy efficiency on Jetson AGX Orin compared to GPGPU-based approaches. These results indicate that IMAX3 can significantly improve both vector and graph-based computations, contributing to the broader acceleration of similarity search systems.