Design of an Energy-Efficient and Reconfigurable Architecture for Accelerating Structurally Sparse Spiking Neural Networks

Name Mingyang Li

Laboratory's name Computing Architecture

Supervisor's name Professor Yasuhiko Nakashima

Abstract (should be within 1st page)

The rapid growth of Internet of Things (IoT) applications has heightened the need for energy-efficient and cost-effective edge computing platforms. However, modern AI workloads, particularly those based on deep learning and convolutional neural networks (CNNs), demand intensive computation and high memory bandwidth, leading to significant overhead and power consumption that challenge deployment on resource-constrained edge devices. Since Von Neumann's architecture suffers from the memory wall, architecture alone cannot solve this issue. In this context, neuromorphic computing and spiking neural networks (SNNs) have emerged as promising alternatives for edge intelligence.

This dissertation proposes FPSpike, a fully-parallel, multi-granularity, multi-hierarchical, and reconfigurable hardware architecture designed to accelerate structured sparse SNNs. Implemented on FPGA, FPSpike leverages structured sparse synaptic connections to reduce neuron computation and memory access costs, while alleviating hardware routing complexity. Its localized interconnects support flexible network partitioning into independent hardware cores, enabling efficient multi-task spatial parallelism without resource redundancy.

To enhance performance, this work proposes a set of architecture-algorithm co-design innovations: (1) cross-layer skip connections mitigate accuracy degradation from structured sparsity and locality constraints, and a task-aware sub-network generation framework enables adaptability to diverse classification tasks; (2) a two-level hierarchical design maximizes on-chip data reuse, supported by programmable processing elements for various spike-based operations. A tailored mapping and dataflow strategy further ensures efficient hardware deployment. FPGA-based experiments show that FPSpike delivers 2.5×–43.9× and 3.2×–93.8× higher inference throughput than CPU and GPU baselines, respectively, demonstrating its strong performance and energy efficiency for next-generation edge intelligence.