Efficient Neural Network Implementation for Embedded Devices in Resource-Constrained Environments

Name: Babak Golbabaei

Laboratory's name: Computing Architecture

Supervisor's name: Yasuhiko Nakashima

As machine learning models continue to be deployed on edge and embedded systems, the challenge of maintaining high performance under strict power, memory, and hardware constraints becomes increasingly important. This dissertation addresses that challenge by proposing novel algorithmic and architectural methods to enable efficient neural network inference in resource-constrained environments.

We introduce a new class of hardware-friendly binary neural networks (BNNs) tailored for embedded platforms. Unlike conventional BNNs that use bipolar representations (−1/+1) and XNOR-based operations, our method adopts a unipolar binary representation (0/1) and utilizes AND-based computation. This shift further simplifies logic, reduces power consumption, and minimizes hardware complexity. To improve accumulation efficiency, we replace population counters with a custom pipelined adder tree structure, reducing the critical path and enhancing timing performance on FPGA implementations.

To address real-world deployment challenges such as noisy or imperfect sensor data, we propose a stochastic input encoding mechanism. This probabilistic approach increases robustness to input variation, enhancing inference stability under noisy conditions. Additionally, we implement a trainable thresholding mechanism inspired by biological neurons, which promotes sparsity and eliminates redundant computation.

Our models are validated on standard datasets such as CIFAR-10, Fashion-MNIST, and MNIST. Despite the architectural simplicity, the proposed networks achieve competitive classification accuracy while significantly lowering power and area usage. FPGA-based implementation confirms real-time inference capabilities with minimal energy consumption. We also explore scaling factor optimization to boost accuracy when necessary and propose parameter encoding schemes to reduce memory overhead.

This work presents a practical and scalable framework for deploying robust and energy-efficient neural networks in real-world, resource-limited scenarios such as IoT, mobile robotics, and wearable devices, particularly where noise tolerance, simplicity, and efficiency are essential.