An Energy Efficient RISC-V based SoC Accelerator For CNN Inference

LUND UNIVERSITY

BY ARTURO PRIETO, MASOUD NOURIPAYAM, KRISTOFFER PER WESTRING, ALEX ALLFJORD, SERGIO CASTILLO MOHEDANO, LORENZ WALTHER, AND JOACHIM RODRIGUES

Department of Electrical and Information Technology

Introduction

Deep Learning (DL) algorithms, like Convolutional Neural Networks (CNN), require extensive computation, making them less suitable for traditional architectures in resource-constrained edge environments. Application-Specific Integrated Circuits (ASICs) offer efficiency but lack adaptability, while Field Programmable Gate Arrays (FPGAs) are flexible but less efficient and more costly. RISC-V processors with customizable co-processors present a promising solution, combining the flexibility of general-purpose processors with the performance of specialized hardware accelerators. This study demonstrates the integration of a radar sensor with a RISC-V platform and co-processor for real-time hand movement detection, showcasing the potential of this approach.

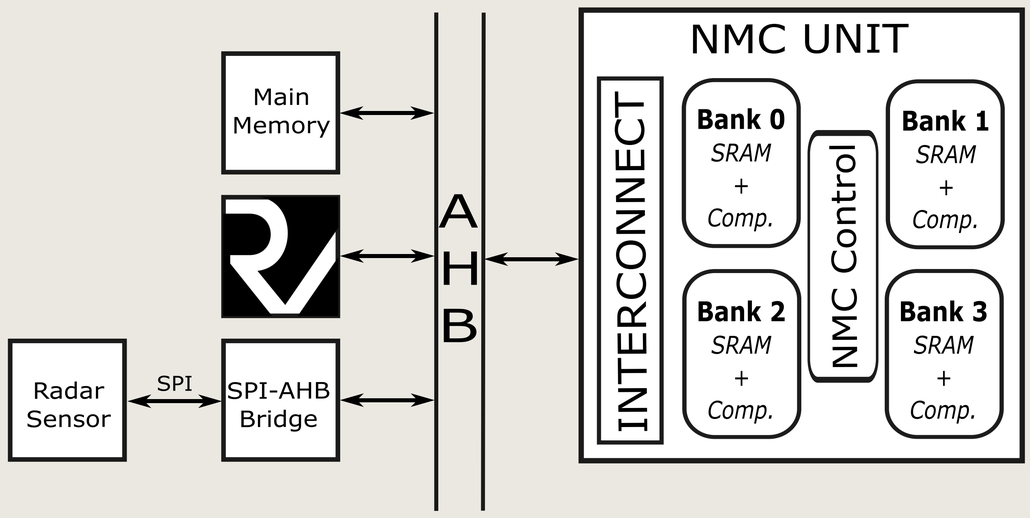
Methodology

This collaborative initiative involves educational institutions and private companies, with private entities providing a state-of-the-art radar sensor and a foundational single-core RISC-V processor. An integrated development environment (IDE) was used for designing and optimizing processor cores, emphasizing the seamless integration and deployment of various accelerator specifications. The IDE facilitated easy modifications to the instruction set, micro-architecture, and peripheral devices, simplifying the incorporation of changes. The system, successfully implemented and tested on an FPGA, demonstrates streamlined processes for both FPGA and ASIC deployments.

System Description

Research efforts are aimed at designing a configurable co-processor architecture optimized for CNN inference, integrated with on-chip memory shared with the RISC-V processor. This setup allows dynamic optimization of power and computational resources, minimizing power usage during lighter workloads and maximizing parallel computation during heavier ones. Additionally, the implementation workflow for ASIC and FPGA-based SoCs is automated, simplifying micro-architectural modifications and diverse CNN implementations. The system uses Block Random Access Memory (BRAM) for main memory and a shared memory space in the co-processor for efficient data processing. Communication between the co-processor and RISC-V is managed via an Advanced High-Performance Bus (AHB), with data streams from the radar sensor handled at the SPI-AHB bridge for translation and preprocessing.

Case Study: Hand Gesture Recognition



The system is evaluated by analyzing the data that is supplied by the radar sensor unit. This data is retrieved from the sensor and undergoes preprocessing before being transferred to shared memory. Subsequently, the co-processor executes concurrently two inferences on two different CNNs, resulting in classification. The quantized pre-trained CNNs enables the system to detect the presence of a hand over the sensor and determine whether the hand is in motion or stationary.

sergio.castillo_mohedano@eit.lth.se

Lund, Sweden