Department of Electronics and Electrical Engineering

भारतीय प्रौद्योगिकी संस्थान- गुवाहाटी
Indian Institute of Technology Guwahati, India

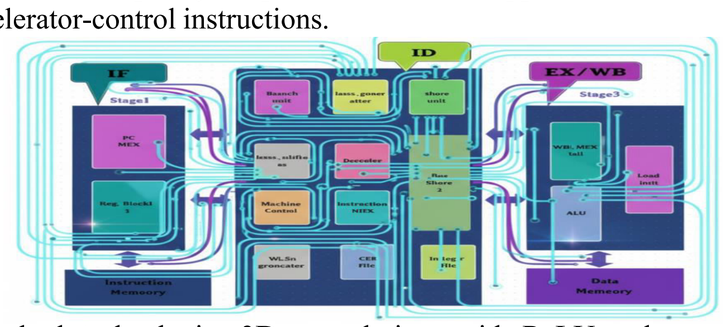
Hardware-Efficient RISC-V Processor and 3D CNN Accelerator with Custom Instruction Support for Edge-AI

Rabisankar Maity: Roll No:244102409
Under the Guidance of Prof. Roy Paily Palathinkal

Abstract: This work presents a hardware-efficient edge-AI platform integrating a 3-stage RISC-V (RV32I) processor with a 3D CNN accelerator and a fixed-point ANN module for handwritten digit recognition. The CNN engine supports standard and depthwise convolutions with ReLU and max-pooling, while optimized BRAM-based memory reuse enables low-latency, high-throughput inference. The RISC-V core, built entirely in Verilog, issues custom instructions to control the accelerator, ensuring tight hardware–software co-design. Functional verification using RTL simulations and Python/OpenCV golden models confirms correctness and robustness. The unified architecture demonstrates scalable, energy-efficient performance for real-time embedded vision and edge-intelligence applications.

1. INTRODUCTION

- This work presents a complete RTL-based exploration of hardware-efficient architectures for next-generation embedded and edge-intelligent systems, integrating custom-designed units aimed at high throughput, low power consumption, and scalable AI-driven workloads.
- A lightweight 3-stage pipelined RISC-V processor is developed with full RV32I compliance and RV32IM extensions, featuring well-organized instruction flow, balanced hardware simplicity vs. performance, and support for enhanced arithmetic and custom accelerator-control instructions.



- Supports standard + depthwise 3D convolutions with ReLU and max-pooling, using FSM control, parallel MAC units, and BRAM-optimized memory reuse to improve throughput and efficiency.
- Implements a multi-layer ANN verified on handwritten digit recognition with a full training-to-hardware workflow, optimized BRAM mapping, and RTL + Python/OpenCV validation
- Combines RISC-V, CNN, and ANN into a unified, reconfigurable AI hardware platform for real-time edge intelligence.

2. MOTIVATION AND PROBLEM FORMULATION

- **Energy- and Memory-Efficient Edge AI Computing:**
  Edge devices require compact and low-power hardware capable of executing complex neural network tasks locally. Conventional CPUs and GPUs are inefficient due to high energy consumption and excessive memory traffic, making them unsuitable for resource-constrained edge environments.
- **Lack of Unified Accelerator Support:**
  Existing hardware accelerators are typically designed for either standard convolution or depthwise convolution, limiting flexibility. This separation leads to underutilization of hardware resources and prevents seamless execution of diverse CNN workloads within a single architecture.
- **Redundant Multi-Modal Processing:**
  Treating CNN and ANN tasks as independent processes increases latency and hardware redundancy in multi-modal AI applications, highlighting the need for a unified platform

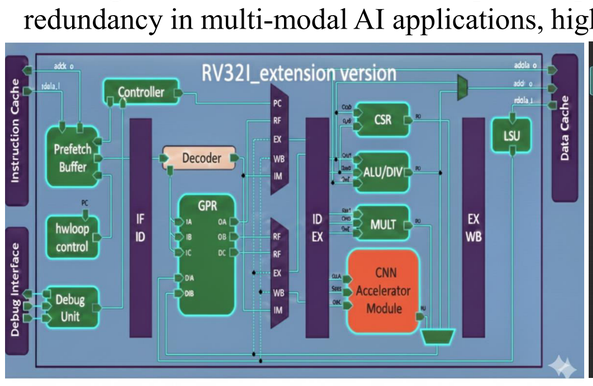


Figure 2(a) : RISC_V_Extension version

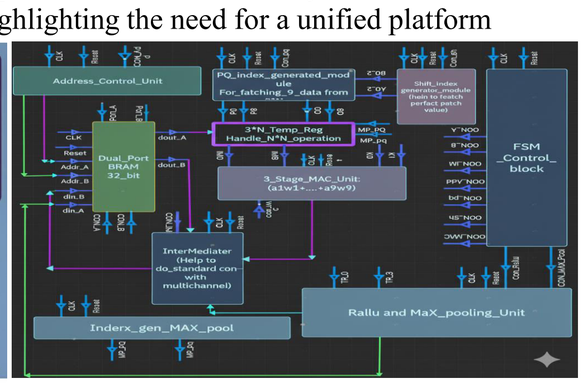


Figure 2(b) : Basic building block of the CNN

- **Integration and Programmability Challenges:**
  The weak integration between processor cores and neural accelerators restricts real-time control, programmability, and adaptability in embedded AI systems. A tightly coupled processor-accelerator architecture is essential for flexible and responsive edge inference.
- **Motivation for a Unified RTL-Based Platform:**
  The proposed fully RTL design integrates a 3-stage RISC-V processor, a multi-channel 3D CNN accelerator with fused ReLU and Max-Pooling, and an optimized ANN for handwritten digit recognition, enabling low-latency, high-throughput, and energy-efficient edge-AI inference.

3. RESULTS AND DISCUSSIONS

- **RISC-V Processor Verification:**
  The 3-stage pipelined RV32I processor was successfully verified using 318 assembly instructions covering arithmetic, logical, load/store, and branch operations. Simulation waveforms confirm correct operation across all pipeline stages (IF, ID, EX/WB), ensuring functional correctness.
- **CNN Accelerator Performance:**
  The standard convolution (256×256×3 input with 3×3×3 kernel, 1 filter) completes in **1.96 ms**, demonstrating efficient pipelined computation and fast feature extraction suitable for real-time edge applications.

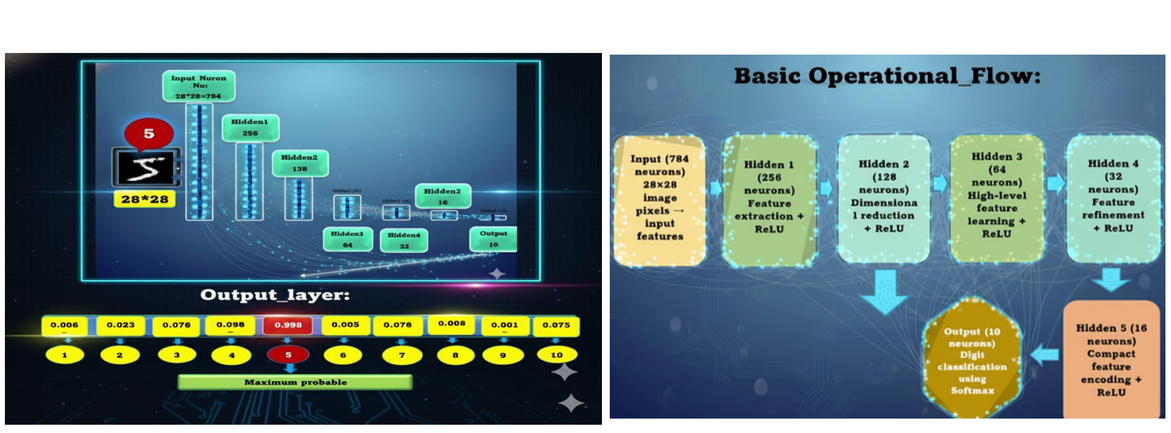
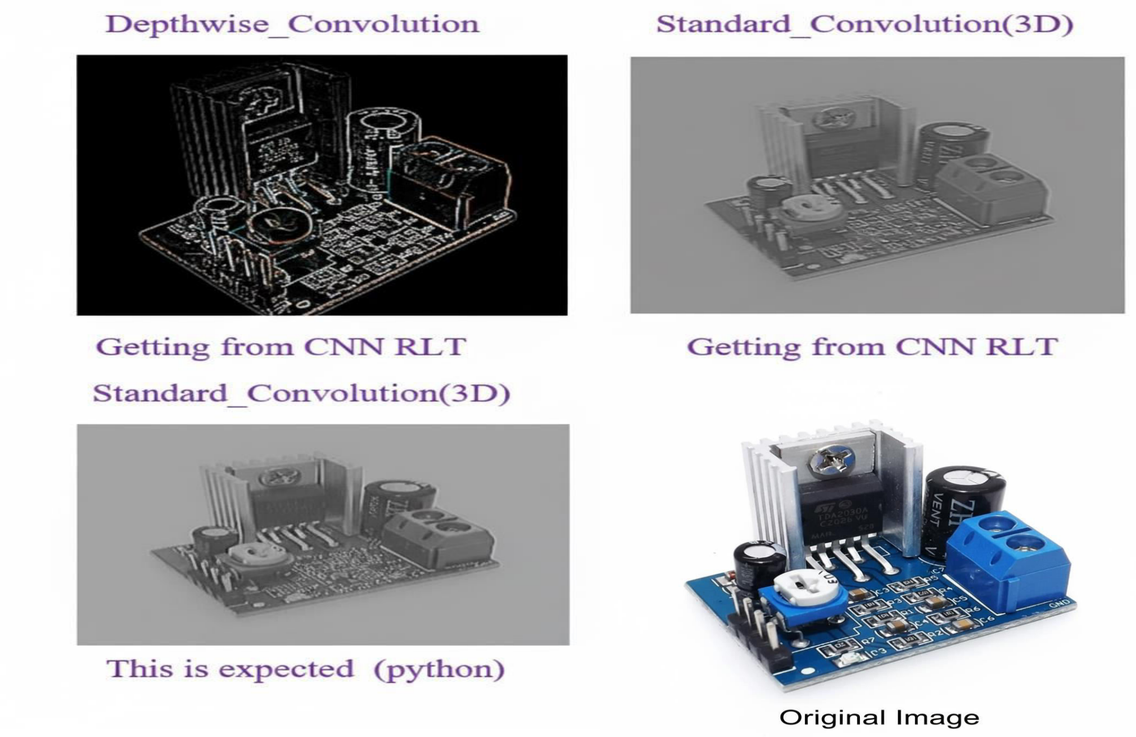


Figure 3(a): Operationsal Flow of ANN



Hardware Utilization Table				
Module / Input Size	LUTs	FFs	BRAM	Notes
RISC-V Processor	47,383	1,713	50	Verified with 318 instruction simulation
CNN 256×256×P (32-bit)	95,729	26,731	256	Channel-independent, optimized reuse
CNN 256×256 (16-bit)	49,338	14,546	256	Reduced bit-width, lower LUT/FF usage
CNN 512×512×P (16-bit)	96,820	26,546	256	Larger image, similar BRAM usage
ANN (Fully Connected)	19,456	26,057	188	Optimized for handwritten digit recognition

Table 1 : Hardware Utilizetion Report of RISC_V CNN ANN

- **ANN Accuracy & Performance:**
  The ANN, configured with extracted weights and biases, classifies handwritten digits (0–9) in **0.794 ms**, demonstrating real-time, high-confidence inference.
- **Resource Efficiency:**
  The design efficiently balances throughput and memory usage, optimizing LUTs, FFs, BRAM, and DSPs for both CNN and ANN modules.

4. FUTURE WORK

- **CNN & RISC-V Optimization:** Accelerate CNN using 16-bit precision and AI-specific RISC-V instructions (MAC, MSUB, CONV, POOL).
- **Model & Edge-AI Applications:** Use quantization and compression for energy-efficient inference and enable real-time applications like gesture recognition.
- **ASIC Implementation:** Transition the verified RTL design to ASIC using standard-cell libraries, focusing on optimizing power, area, and timing while ensuring high-performance, reliable, and silicon-efficient realization for edge-AI applications.

5. REFERENCES

- 1] S. Nair, A. Chatterjee, and K. K. Parhi, “RISC-V Based Custom Instruction Set Extension for CNN Acceleration on Edge Devices,” IEEE Access, vol. 11, pp. 12345–12356, 2023.
- [2] Y. Zhang, M. Chen, and H. Lee, “Design and Implementation of a Fused RISC-V and CNN Co-Processor for Edge AI Applications,” Microprocessors and Microsystems, vol. 101, pp. 1–12, 2024.
- [3] H. Xu and C. Shen, “Winograd and IM2COL-Based Optimization for Low-Power CNN Accelerators,” IEEE Trans. VLSI Syst., vol. 31, no. 4, pp. 512–523, 2023.