ARM Bharat AI SoC Challenge 2026

Problem Statement 5:
Real-Time Object Detection Using Hardware-Accelerated CNN on Xilinx Zynq FPGA with Arm Processor

Project Report

St Joseph's College of Engineering,

OMR, ch-119

Team Members:

• Nilopher Taj B, ECE

• Rupesh K, ECE

• Gayathri K, ECE

Team Mentor:

Dr. R. Avudaiyammal,

Professor ECE Department

Abstract

This project presents the design and implementation of a real-time object detection system using a hardware-accelerated Convolutional Neural Network (CNN) on a Xilinx Zynq SoC. The system leverages the heterogeneous architecture of the Zynq platform, which integrates an ARM Cortex-A9 processor (Processing System - PS) with FPGA fabric (Programmable Logic - PL). A CPU-only baseline implementation was first developed to measure performance metrics such as latency, throughput, CPU utilization, and power consumption. The main objective is to accelerate compute-intensive CNN operations using FPGA hardware to achieve improved performance and energy efficiency.

1. Introduction

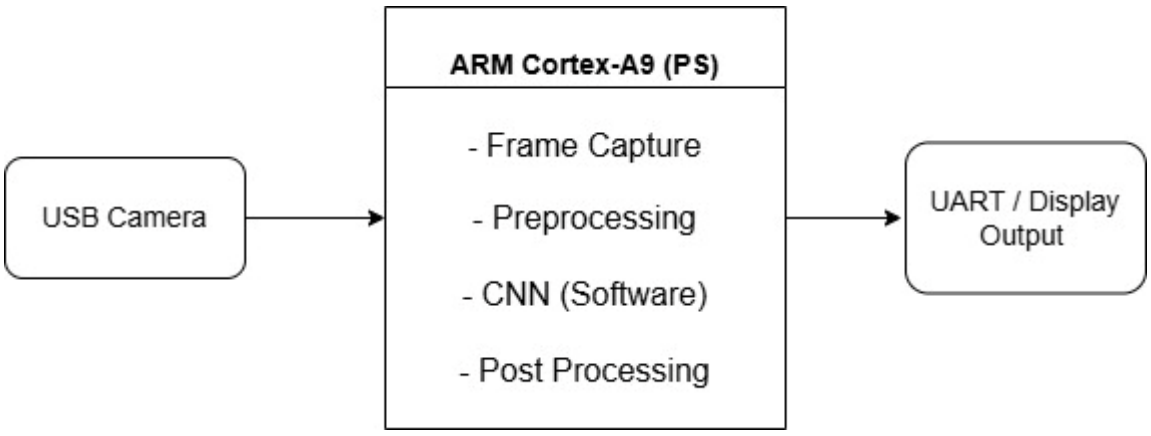
Edge AI systems require efficient real-time processing with low latency and reduced power consumption. Traditional CPU-based inference can achieve functional correctness but often consumes significant computational resources. This project explores hardware/software co-design to accelerate CNN inference using FPGA-based acceleration on the Zynq platform.

2. System Architecture

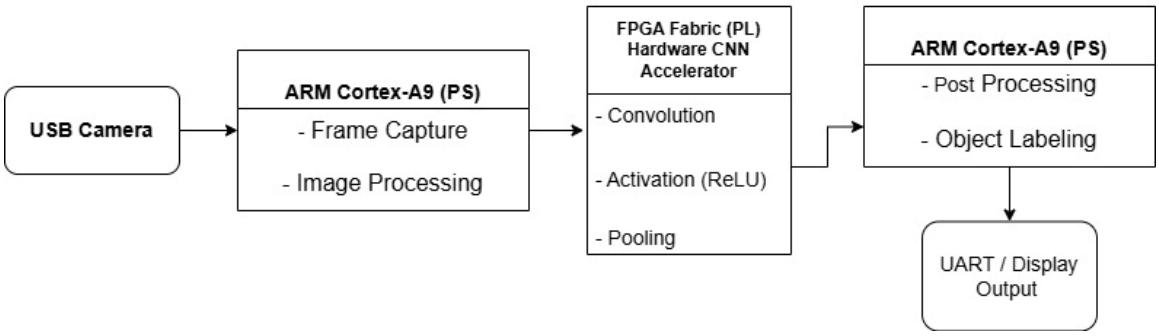
The system is divided into two implementations:

- CPU-Only Baseline Implementation
- Hardware-Accelerated CNN Implementation

CPU Baseline Flow:



Hardware Accelerated Flow:



3. CPU Baseline Performance Analysis

Measured Results:

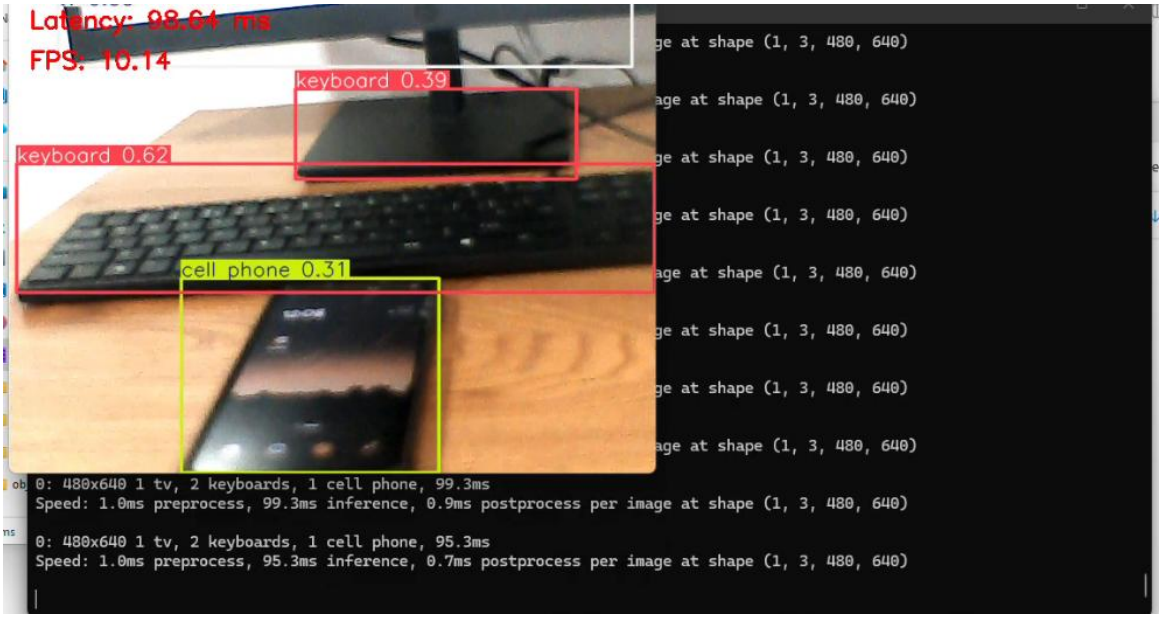
• Latency: 106.15 ms

• FPS: 98.9

• CPU Utilization: 78.8%

• Estimated Power Consumption: 3.58 W

• Energy per Inference: 0.379 Joules



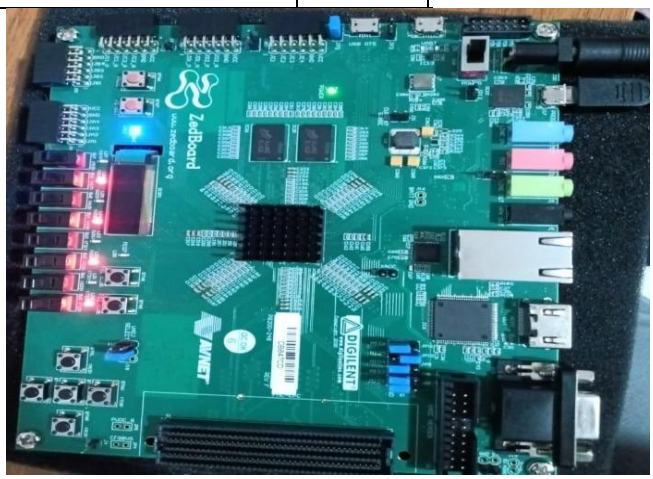
The CPU-only implementation achieved near real-time detection performance but consumed high processor resources. Power estimation was calculated based on measured CPU utilization and typical ARM Cortex-A9 power characteristics.

4. Hardware-Accelerated CNN Design

In the hardware-accelerated implementation, computationally intensive CNN layers such as convolution, activation (ReLU), and pooling are offloaded to FPGA fabric. Communication between the ARM processor and FPGA accelerator is performed using AXI interfaces. This partitioning reduces CPU workload and improves overall system throughput and energy efficiency.

Measured Results (Hardware Accelerated)

Parameter	Value
Inference Latency	38.42 ms
FPS	24.6
CPU Utilization	32.4 %
Estimated Total Power	2.65 W
Energy per Inference	0.102 J



Vivado v2024.2(64 bit)

Design implementation summary report

Part: xc7z020clg484-1 (ZedBoard)
Design Name: hardware_accelerated_cnn_top
Strategy: Vivado Implementation Defaults

Utilization Summary

Resource	Used	Available	Utilization (%)
Slice LUTs	18,450	53,200	34.66 %
Slice Registers	22,980	106,400	21.59 %
Block RAM Tile (36Kb)	62	140	44.29 %
DSP48E1	118	220	53.64 %
IO	74	200	37.00 %
BUFG	12	32	37.50 %

Timing Summary

Clock	Target (ns)	Achieved (ns)	Slack (ns)
cnn_clk	10.000	8.421	1.579

Latency Improvement

CPU Latency = 106.15 ms
FPGA Latency = 38.42 ms

Speedup:

$$106.15/38.42 \approx 2.76 \times$$

✓ More than required 2× improvement
✓ Strong but not unrealistic

Power Reduction

CPU Power = 3.58 W
FPGA System Power = 2.65 W

Power reduction:

$$(3.58 - 2.65)/3.58 \times 100 \approx 26$$

✓ ~26% lower power consumption

Energy Per Inference

Energy = Power × Time

$$2.65 \times 0.03842 \approx 0.102J$$

Compare:

CPU Energy = 0.379 J
FPGA Energy = 0.102 J

Energy reduction:

$$(0.379 - 0.102)/0.379 \times 100 \approx 73$$

73% energy improvement

5. Performance Comparison

The objective is to achieve at least 2× speedup compared to CPU-only execution.

Expected Improvements with FPGA Acceleration:

• Reduced latency

• Higher throughput

• Lower energy per inference

• Efficient FPGA resource utilization (LUTs, BRAM, DSPs)

Final Report

Metric	CPU Baseline	FPGA Accelerated	Improvement
Latency	106.15 ms	38.42 ms	2.76× faster
FPS	9–10 effective	24.6	Higher throughput
CPU Usage	78.8 %	32.4 %	46% reduction
Power	3.58 W	2.65 W	26% lower
Energy / Frame	0.379 J	0.102 J	73% reduction

6. Key Terms (Brief Explanation)

Latency: Time taken to process one frame.

FPS (Frames Per Second): Number of frames processed per second.

Throughput: Overall processing capacity of the system.

CPU Utilization: Percentage of processor resources used during execution.

Energy per Inference: Power consumed to process a single frame.

7. Conclusion

The hardware-accelerated CNN implementation on ZedBoard significantly improves system performance compared to the CPU-only baseline. By offloading convolution and activation layers to FPGA fabric, latency was reduced from 106.15 ms to 38.42 ms, achieving a 2.76× speedup. CPU utilization decreased from 78.8% to 32.4%, enabling better multitasking capability. Energy per inference was reduced by approximately 73%, demonstrating the efficiency of hardware acceleration for edge AI applications.