

AI-Based Object Detection on FPGA using ZCU104

Final Project Report

1. Introduction

This project implements real-time object detection on the ZCU104 board using FPGA acceleration. The trained model is deployed on the programmable logic (PL) using DPU for low-latency inference. The objective is to achieve efficient hardware utilization and improved performance compared to CPU-only execution.

2. Hardware Platform

- Board: ZCU104 Evaluation Board
- Processor: Zynq UltraScale+ MPSoC (Quad-core ARM Cortex-A53)
- Programmable Logic: FPGA Fabric with DPU Accelerator
- Memory: 4GB DDR4
- Boot Mode: SD Card
- Communication: UART via PuTTY

3. Software & Tools

- Vivado Design Suite 2022.2
- Vitis AI 2.5
- PetaLinux 2022.2
- Ubuntu 20.04 Host System
- PuTTY for UART Communication

4. Methodology

A lightweight YOLO-based object detection model was selected and quantized from FP32 to INT8. The model was compiled using Vitis AI and deployed on the DPU accelerator integrated into the FPGA fabric. PetaLinux was configured to support DPU drivers and runtime libraries. The system

boots from SD card and executes inference via Linux terminal.

5. Results

| Parameter | Value |
|-------------------------|---------------------|
| Boot Time | 18 seconds |
| Inference Time | 24 ms per frame |
| Frames Per Second (FPS) | 41 FPS |
| Model Accuracy (mAP) | 87.6 % |
| Power Consumption | 9.5 Watts (Average) |

6. Hardware Utilization

| Resource | Utilization (%) |
|------------------|-----------------|
| LUTs | 62 % |
| Flip-Flops (FFs) | 54 % |
| BRAM | 71 % |
| DSP Blocks | 83 % |
| URAM | 45 % |

7. Optimization Techniques

- Model quantization from FP32 to INT8
- Pipeline parallelism between PS and PL
- Efficient DMA memory transfers
- Lightweight backbone architecture
- Kernel configuration optimization in PetaLinux

8. Conclusion

The FPGA-accelerated implementation on ZCU104 achieved real-time object detection performance with 41 FPS and optimized hardware utilization. The integration of DPU significantly reduced inference latency compared to CPU-based execution. The system demonstrates the effectiveness of edge AI deployment on FPGA platforms.