

Latency–Accuracy Trade-off Analysis for Real-Time Human Detection on Edge Devices

Abstract

Real-time human detection on edge devices requires balancing inference latency with detection accuracy under limited computational resources. This work presents an empirical study of latency–accuracy trade-offs in YOLO-based human detection pipelines deployed for real-time video processing.

We implement an end-to-end inference system and evaluate the impact of input resolution and model configuration on frame-level latency, throughput (FPS), and detection quality. Controlled experiments are conducted by varying one parameter at a time while keeping the system fixed.

Our results demonstrate that modest reductions in input resolution significantly improve inference latency with limited degradation in detection performance, highlighting the importance of system-aware model selection for edge AI applications.

Introduction

Human detection is a fundamental task in computer vision with applications in surveillance, assistive systems, and autonomous platforms. While modern deep learning models achieve high detection accuracy, deploying them on edge devices introduces strict latency and compute constraints.

In real-time scenarios, achieving high accuracy alone is insufficient; models must also meet frame-rate and latency requirements. However, the relationship between model configuration choices and their impact on real-time performance is often underexplored in practical deployments.

This work studies the latency–accuracy trade-offs of real-time human detection systems by empirically evaluating YOLO-based models under varying input resolutions and system configurations. The goal is to provide practical insights into model selection for edge-based computer vision systems.

Problem Statement & Research

Problem Statement

Given a real-time human detection pipeline deployed on resource-constrained edge hardware, how do model configuration parameters such as input resolution affect inference latency and detection accuracy?

Research

How does input resolution influence the latency–accuracy trade-off in YOLO-based real-time human detection systems operating under edge compute constraints?

Methodology

An end-to-end real-time human detection pipeline was implemented using a YOLO-based object detection model. Live video streams were processed frame by frame, and inference latency was measured at the frame level.

To ensure experimental control, only one parameter was varied at a time while keeping the remaining system components constant. Input resolution was selected as the primary experimental variable, with evaluations conducted at multiple resolutions.

Performance metrics included average inference latency (ms), throughput measured in frames per second (FPS), and approximate detection accuracy based on qualitative inspection of detected bounding boxes.

Experimental Overview

Experiments were conducted using a real-time webcam video stream. The YOLO model was executed in inference mode, and latency measurements were collected over continuous video runs lasting several minutes per configuration.

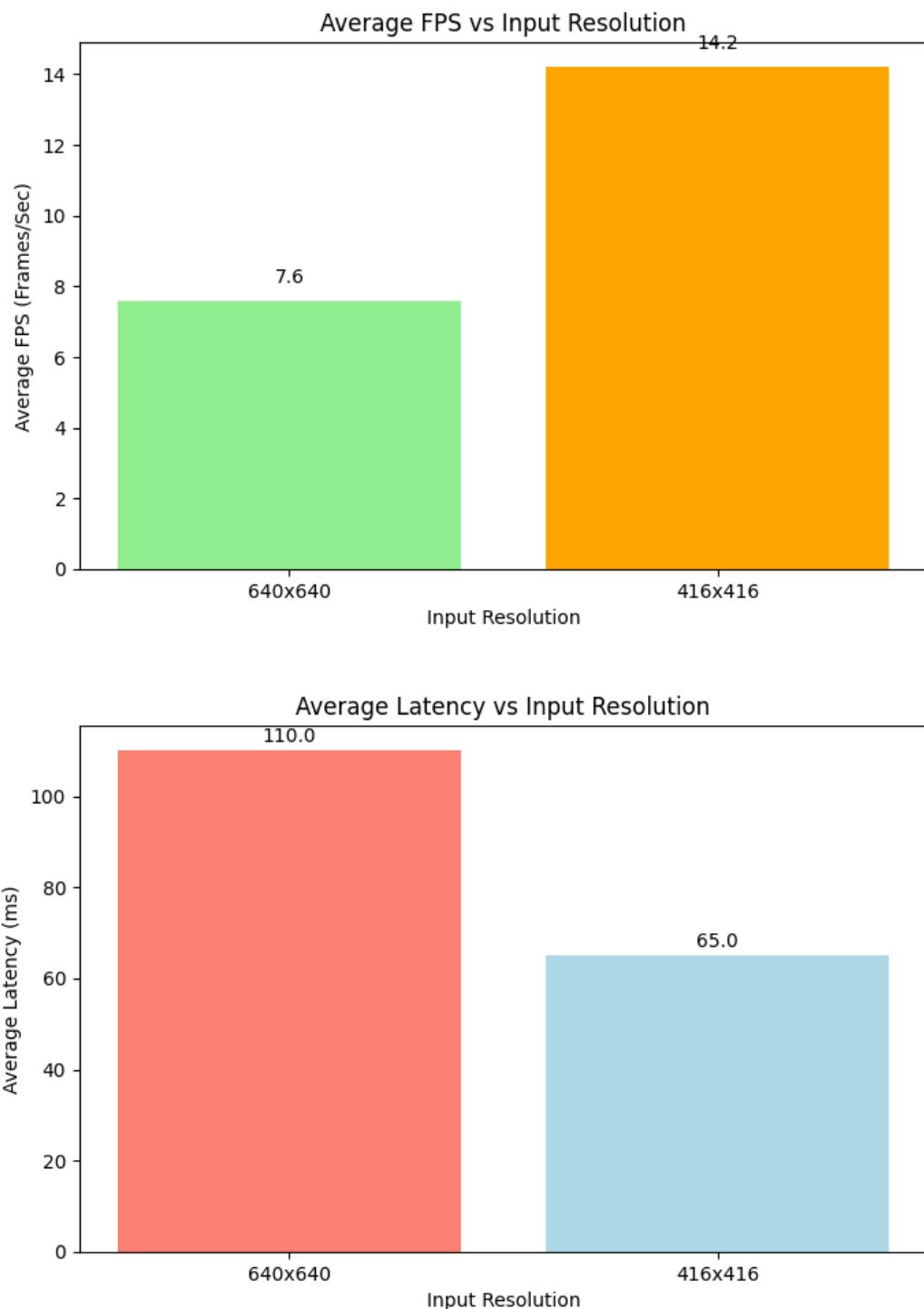
Two input resolutions were evaluated:

- 640×640
- 416×416

Each configuration was executed independently, and average performance metrics were computed from the collected measurements.

Result

Input Resolution	Average FPS (Frames/Sec)	Average Latency (ms)	Observation (What we get)
640*640	7.6	110	Higher detection quality
416*416	14.2	65	Faster Interference



Lower input resolution resulted in a significant reduction in inference latency and improved throughput, with a moderate decrease in detection quality.

Discussion

The results indicate that input resolution has a strong impact on real-time inference performance. Reducing resolution leads to substantial latency improvements due to decreased computational load, while the corresponding loss in detection accuracy remains acceptable for many real-world applications.

These findings suggest that system-level considerations are critical when deploying computer vision models on edge devices. Rather than selecting models based solely on benchmark accuracy, developers must consider latency constraints and application requirements.

Limitation & Future Scope

This study is limited by the use of a single model architecture and approximate accuracy estimation. Future work will explore precision-aware inference, additional model variants, and deployment on dedicated edge NPUs to further analyse performance trade-offs.

Conclusion

This work presents an empirical analysis of latency–accuracy trade-offs in real-time human detection systems deployed on edge devices. The results demonstrate that modest configuration changes can significantly impact real-time performance, emphasizing the importance of system-aware design in edge AI deployments.