

INTISARI

Deteksi objek secara *real-time* pada perangkat seperti CCTV menghadapi dua tantangan utama: penurunan akurasi yang signifikan dalam kondisi pencahayaan minim dan keterbatasan sumber daya komputasi perangkat keras. Model deep learning seperti YOLOv5, yang umumnya beroperasi pada presisi *floating-point* 32-bit (FP32), memiliki ukuran file yang besar dan latensi inferensi yang tinggi, sehingga tidak efisien untuk implementasi *on-device*. Kondisi minim cahaya, yang dipenuhi noise dan kontras rendah, semakin menurunkan kinerja deteksi model-model ini. Penelitian ini bertujuan mengevaluasi efisiensi kinerja YOLOv5 menggunakan teknik *Post-Training Quantization* (PTQ) INT8 untuk deteksi objek secara real-time pada kondisi pencahayaan minim.

Eksperimen dilakukan menggunakan aliran video CCTV secara langsung selama 60 menit pada perangkat laptop berbasis prosesor AMD tanpa GPU khusus. Model YOLOv5 FP32 dibandingkan dengan model YOLOv5 yang dikonversi ke ONNX dan dikuantisasi menjadi INT8 dengan preprocessing CLAHE yang sama. Hasil pengujian menunjukkan bahwa model YOLOv5 FP32 memperoleh akurasi rata-rata 44,89% dengan 5,65 FPS, penggunaan CPU 60,95%, dan RAM 89,85%, sedangkan model ONNX INT8 memperoleh akurasi 55,27% dengan 8,60 FPS, penggunaan CPU 52,05%, dan RAM 85,35%. Hasil tersebut menunjukkan bahwa pada skenario pengujian penelitian ini, penerapan PTQ INT8 meningkatkan kecepatan inferensi dan menurunkan penggunaan sumber daya komputasi, serta menghasilkan nilai akurasi yang lebih tinggi dibandingkan model FP32.

Kata kunci: Deteksi Objek, YOLOv5, Pencahayaan Minim, ONNX, PTQ

ABSTRACT

Real-time object detection on devices such as CCTV cameras faces two primary challenges: a significant decrease in accuracy under low-light conditions and the limited computational resources of the hardware. Deep learning models like YOLOv5, which typically operate at 32-bit floating-point (FP32) precision, suffer from large file sizes and high inference latency, making them inefficient for on-device deployment. Low-light environments, characterized by high noise and low contrast, further degrade the detection performance of these models. This study aims to evaluate the performance efficiency of YOLOv5 utilizing the INT8 Post-Training Quantization (PTQ) technique for real-time object detection in low-light conditions.

The experiment was conducted using a live CCTV video stream for 60 minutes on a laptop equipped with an AMD processor and no dedicated GPU. The baseline YOLOv5 FP32 model was compared against a YOLOv5 model converted to ONNX and quantized to INT8, with both utilizing the same CLAHE preprocessing technique. The test results demonstrate that the YOLOv5 FP32 model achieved an average accuracy of 44.89% at 5.65 FPS, consuming 60.95% of the CPU and 89.85% of the RAM. In contrast, the ONNX INT8 model achieved a higher accuracy of 55.27% at 8.60 FPS, with lower resource consumption at 52.05% CPU and 85.35% RAM usage. These findings indicate that, within the testing scenarios of this study, applying INT8 PTQ not only accelerates inference speed and reduces computation

Keywords: Object Detection, YOLOv5, Quantization, Low-Light, ONNX