NANYANG TECHNOLOGICAL UNIVERSITY
SINGAPORE

School of Computer Science and Engineering
College of Engineering

QuickYOLO

Ultra-Low Power Real-Time Object Detection based on Quantized CNNs

Student: Chew Jing Wei

Supervisor: Asst Prof Liu Weichen

Model	mAP@0.5	Inference Time (ms)	Peak Memory Usage (MB)	Speedup	Memory Use Reduction
YOLOv2	0.42	577.9	261.9	**21.8X**	**15.3X**
QuickYOLO	0.37	26.5	17.1		

Table 1: Experimental results on 224x224 inputs, generated from TFLite benchmarking tool on Jetson Nano using 4 threads.

Project Objectives

- YOLO-based object-detection model, binarized input & weights
- 30 FPS on edge device
- Reasonable accuracy

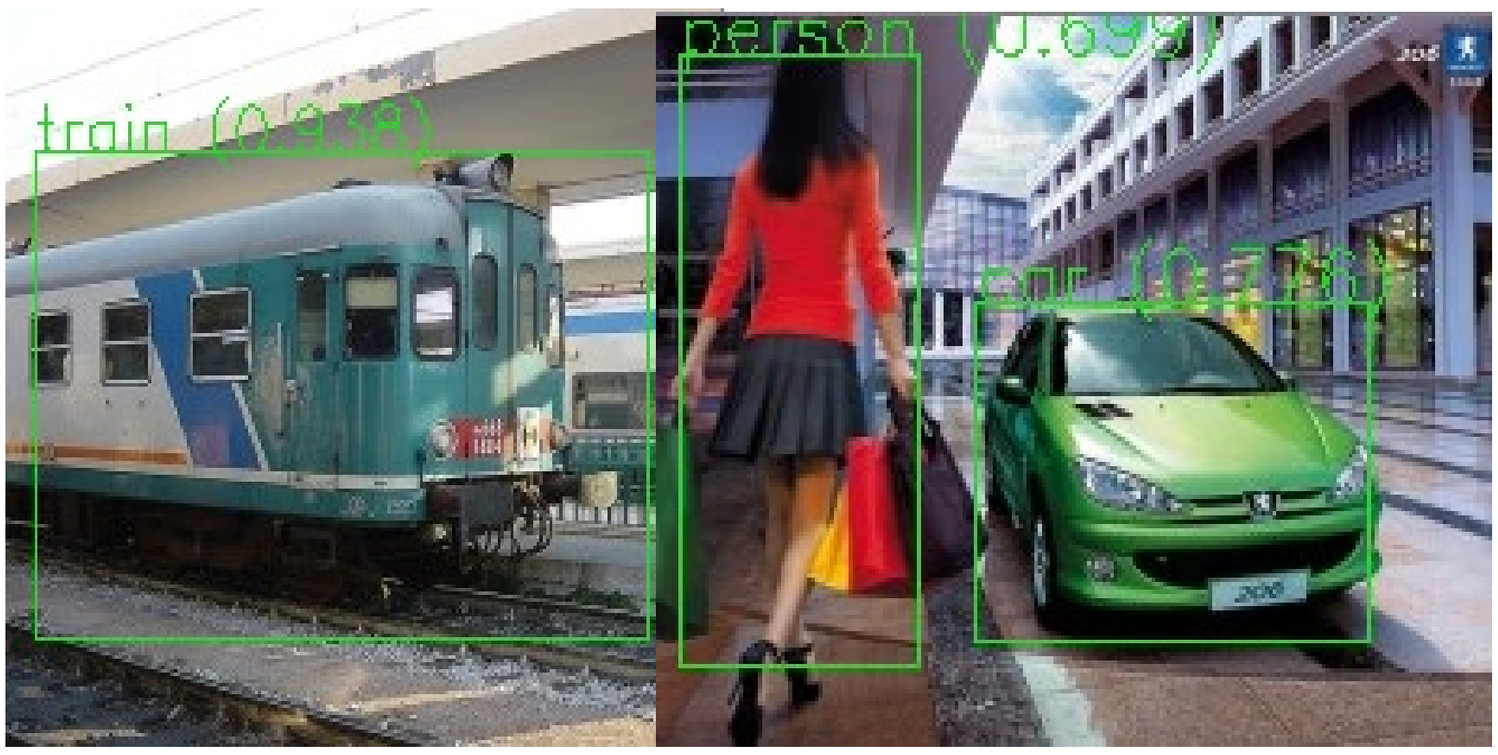

Figure 1: Example predictions from QuickYOLO on the VOC2007 test set

Approach

Find and tweak existing **YOLO** projects in **Python**

Train binarized model using **TensorFlow** & **Larq**

Export using **TensorFlow Lite** & **Larq Compute Engine**

Write a demo program in **C++** with **OpenCV** for camera input