Faster R-CNN Inception V2 Model Performance and Validation Report on Intel® NUC 11 Extreme Kit

Report

August 2023

Document Number: 784004-1.0

Contents

Figures

Tables

Revision History

Date	Revision	Description
August 2023	1.0	Initial release.

§

1.0 Overview

This document provides an overview and performance results for validation of a production ready version of Automatic AI-based Vision Inspection of Vehicles solution with Combi Model (Faster R-CNN Inception V2) by Deevia Software running on an Intel® NUC 11 Extreme Kit with 11th Generation Intel® Core™ Processors.

1.1 Objective

The objective of the validation process is to:

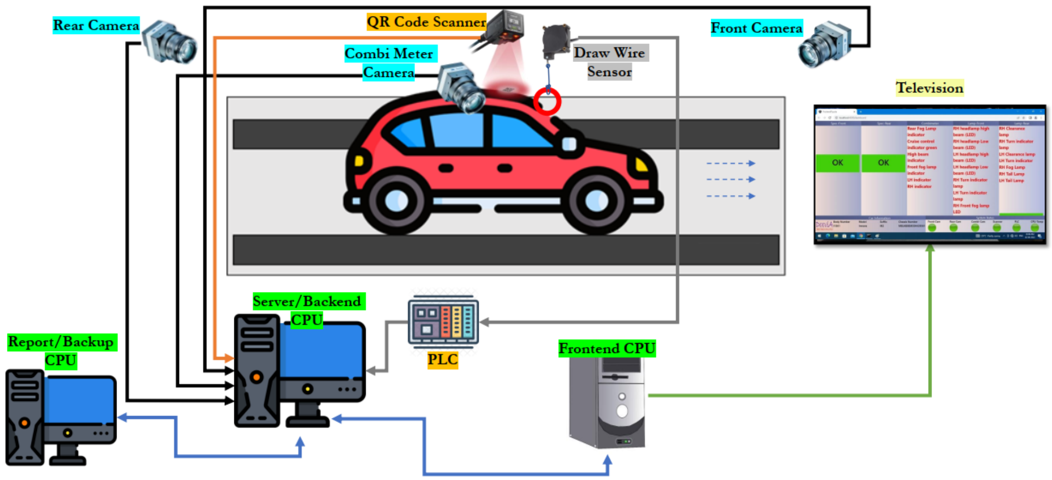
i. Improve inference performance of Combi (Faster R-CNN Inception V2) Topology on Intel® Core™ Processors.
ii. Run benchmarks comparing inference performance on throughput and latency using CPU Baseline versus using the Intel® Arc™ A770 Graphics with Intel® Core™ Processors – Edge
iii. Showcase inference performance gains on throughput with lower latency using Intel® Artificial Intelligence and Deep Learning Solutions software and hardware stack.

1.2 Workload Description

Automatic AI-based Vision Inspection of Toyota's Premium Vehicles with high TAT (<30 sec/vehicle) and Quality (99% Accuracy).

The solution uses state-of-the-art deep learning technologies and image processing algorithms for inspecting assembled vehicles. The solution includes specification checks, lamp functionality checks and dashboard/combi-meter functional verification.

1.3 Automatic AI based Vision Inspection of Vehicles Solution Architecture Overview



Vision based Vehicle Inspection – System Diagram

The following steps provide an overview of the end-to-end Vision based Vehicle Inspection solution control flow:

1. **QR Code Scanning**
   - Inspection Start.
   - Fetch the features to be inspected for the Suffix.
2. **Specification Check (Front and Rear)**
   - AI (ML)-based check for Presence/Absence.
   - Image Processing based Type Check.
   - 5 Seconds (30 Frames) 5 Seconds (30 Frames).
3. **Lamp Inspection (Front and Rear)**
   - Based on Lamp part identification during spec check, region identification for individual lamps.
   - Exposure change of Cameras for Lamp Inspection.
   - Image processing based Functional and Timing checks.
   - 10-20 Seconds (Based on Operations).
4. **Combi Meter Check**
   - AI (ML)-based check for Presence/Absence.
   - Logic based functional and timing checks.
   - 15-25 Seconds (based on Operations).
5. **End Inspection**
   - Inspection Automatically Ends if all the Results are PASS.
   - In case of any FAIL, the Operator pulls the trigger to End the process.
   - System Reset for the Next Vehicle Inspection.

1.4 Intel® Distribution of OpenVINO™ Toolkit

OpenVINO™ is an open-source toolkit for optimizing and deploying AI inference.

1. Boost deep learning performance in computer vision, automatic speech recognition, natural language processing and other common tasks.
2. Use models trained with popular frameworks like TensorFlow, PyTorch and more.
3. Reduce resource demands and efficiently deploy on a range of Intel® platforms from edge to cloud.

Click here to learn more about the Intel® Distribution of OpenVINO™ Toolkit.

§

2.0 System Configuration

Table 1. System Configuration

Components	
Hardware	
Chassis	Intel® NUC 11 Extreme Kit - NUC11BTMi9
CPU	11th Gen Intel® Core™ i9-11900KB @ 3.30GHz
Memory	Installed Physical Memory (RAM) of 32 GB (2x 16GB DDR4 3200 MT/s [3200 MT/s])
Hard drives	512GB Total Storage but not leveraged for storage
Network card	Intel® Ethernet Controller i225-LM
Software	
BIOS	Intel Corp.
Version	DBTGL579.0064.2022.1108.1552
Operating System	Ubuntu 22.04.2 LTS
Kernel	5.19.0-42-generic
Microcode	0x42
IRQ Balance	Enabled
Analytics Application	Faster R-CNN with Inception V2 Model
OpenVINO™	OpenVINO™ Toolkit v2022.3
OpenSSL	OpenSSL 3.0.2 15 Mar 2022 (Library: OpenSSL 3.0.2 15 Mar 2022)
Others	Hyper Threading (Logical Processor in BIOS)
Others	Enabled Dynamic CPU Frequency

2.1 dGPU Details

Manufacturer	Model
Intel	Intel® Arc™ A770 Graphics

2.2 Workload Parameters

Parameters		
Workload		Combi (Faster RCNN Inception V2) model has been shared by the customer.
		Used OMZ Object detection Python demo for the inference. Reference Link: https://docs.openvino.ai/2022.3/omz_demos_object_detection_demo_python.html
Framework	Version	Python 3.10.6, TensorFlow v.1.15.2 and OpenVINO™ Toolkit – 2022.3
Framework Parameters	Warm up steps	Average of 5 runs and the run method was Warm
	Steps	-
	-a flag	-
	e	Did not set no. of inter threads
Model Parameters	Batch Size	1
Dataset	No of Images	7846
	Shape	Customer Dataset, Input shape – [1,540,960,3], Input video streams - 94

3.0 Faster R-CNN Inception V2 Software Configuration

3.1 Configuration

	Settings	Comments
Configuration 1 (baseline)	11th Gen Intel® Core™ i9-11900KB with **CPU** **Configuration**: TensorFlow 1.15.2 OpenVINO™ 2022.3 Python 3.10.6 Input shape - [1,540,960,3]	Running benchmarks comparing inference performance on throughput and latency using CPU Baseline vs Intel® Arc™ A770 dGPU Optimized on Intel® Core™ i9 processor.
Configuration 2 (optimized)	11th Gen Intel® Core™ i9-11900KB with **Intel® Arc™ A770 dGPU** **Configuration**: TensorFlow 1.15.2 OpenVINO™ 2022.3 Python 3.10.6 Input shape - [1,540,960,3]	

§

4.0 Profiling

4.1 Validation Steps

1. Deploy and Configure Intel® NUC.
2. Install Ubuntu* Operating System and Analytics Platform with Testing Criteria.
3. Install OpenVINO™, TensorFlow and the required packages like *numpy*, OpenCV, and so on.
4. Set up the workload on the NUC and convert the TensorFlow model to the Intermediate Representation (IR) model format.
5. Run the workload to Inference the model with OpenVINO™ on CPU and dGPU hardware.
6. Run the profiler tool to record hardware usage and other metrics.
7. Process results to generate tabulated data using multiple readings.
8. Analyze results and report.

4.2 Checklist for Results Validation

i. Processing frame rate is matching the expectations.

ii. Improve the Inference throughput.

iii. Reduce the Latency.

§

5.0 Performance Test Results

CPU	Input
11th Gen Intel® Core™ i9-11900KB @ 3.30GHz	Video (*.avi format). Shape - [1,540,960,3]

5.1 Throughput

Throughput			SW Config 1 Config1 (baseline) OpenVINO™ **11th Gen Intel® Core™ i9-11900KB with CPU only**	SW Config 2 Config2 (Optimized) OpenVINO™ **11th Gen Intel® Core™ i9-11900KB with Intel® Arc™ A770 Graphics**	Speed Up (Gains over baseline) Config 2 vs. Config 1
Model	Precision	Batch size	Throughput (FPS)	Throughput (FPS)	
Combi	**FP32**	1	10.94	62.14	5.68
Combi	**FP16**	1	10.82	32.36	2.99
Combi	**INT8**	1	29.16	32.22	1.1

5.2 Latency

Latency			SW Config 1 Config1 (baseline) OpenVINO™ **11th Gen Intel® Core™ i9-11900KB**	SW Config 2 Config2 (Optimized) OpenVINO™ **11th Gen Intel® Core™ i9-11900KB with Intel® Arc™ A770 Graphics**	Speed Up (Gains over baseline) Config 2 vs. Config 1
Model	Precision	Batch size	Latency (ms)	Latency (ms)	
Combi	**FP32**	1	422.94	37.58	11.25
Combi	**FP16**	1	427.98	70.44	6.07
Combi	**INT8**	1	132.38	70.24	1.88

6.0 Conclusion

Based on the analysis in this report, we have defined the specifications required for CPU and dGPU hardware to be deployed for increasing the Combi model performance using the Intel® NUC 11 with the Intel® Core™ i9-11900KB @ 3.30GHz.

§