Performance and Functionality testing of Intel Neural Compute Stick 2 - Movidius NCS2

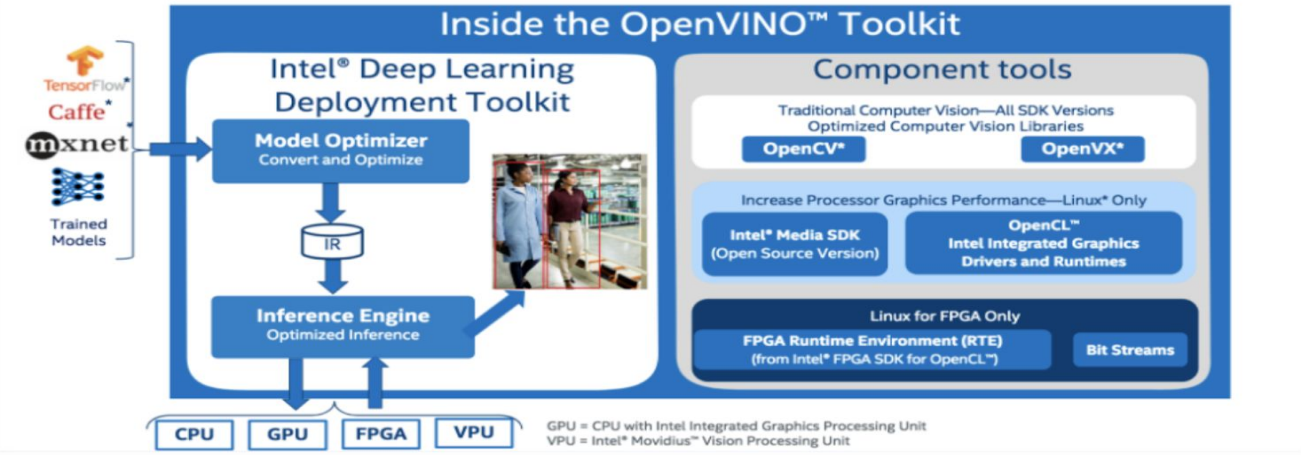
Divija Palleti - September 7th, 2019

Overview

The focus of the following report is to analyse the roll of NCS2 in accelerating deep learning inferencing. Our main focus is on convolution neural networks for images and videos. We make use of the pertained models and OpenVINO tool kit to optimize pre-trained networks and run them on different Intel hardware, such as CPU, and VPU.

OpenVINO Toolkit

The OpenVINO Toolkit is an open source toolkit from Intel. It works well with many pre-trained models in Caffe, TensorFlow or MXNet formats. The Model Optimizer of this toolkit converts the model into an intermediate format (IR) and performs some basic optimizations. The Inference Engine can then run the network on Intel CPUs, GPUsor VPUs (Movidius NCSI2). OpenVINO also contains tools for pre-processing and post-processing data which can be accelerated on CPUs or GPUs. OpenVINO uses a highly optimized library for CPU execution. So utilizing OpenVINO on your model will still provide performance improvements over running them through with Caffe, TensorFlow or MXNet on the same CPU.

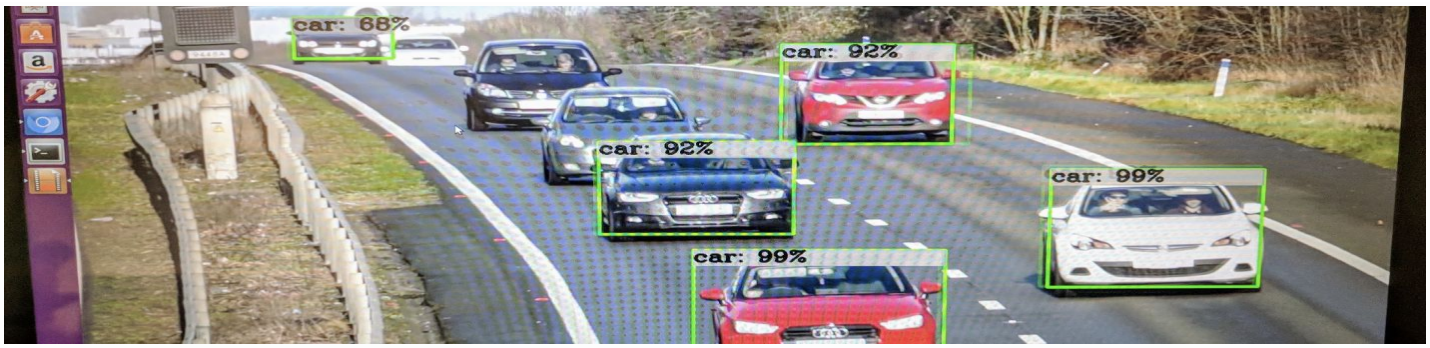


Functionality testing

Inorder to test if the NCS2 is powerful enough to process a real-time video, we used a webcam and fed the video stream to a sample application at http://www.github.com/movidius/ncappzoo . ncappzoo is an open source github repository that contains numerous examples with a simple layout and easy to use Makefiles especially tailored for the Intel® NCS 2 and OpenVINO developer community and helps developers get started quickly by focusing on application code that use pretrained neural networks.

This particular demo app actually predicts the age of the person. In this case NCS2 performs very well, but the accuracy on the age detection isn't at it's best but we are hoping that it is mainly due to the algorithm or training data.

We have also tested NCS2 for object detection with OpenVINO™ toolkit. This application uses the Caffe implementation of the MobileNet SSD model.

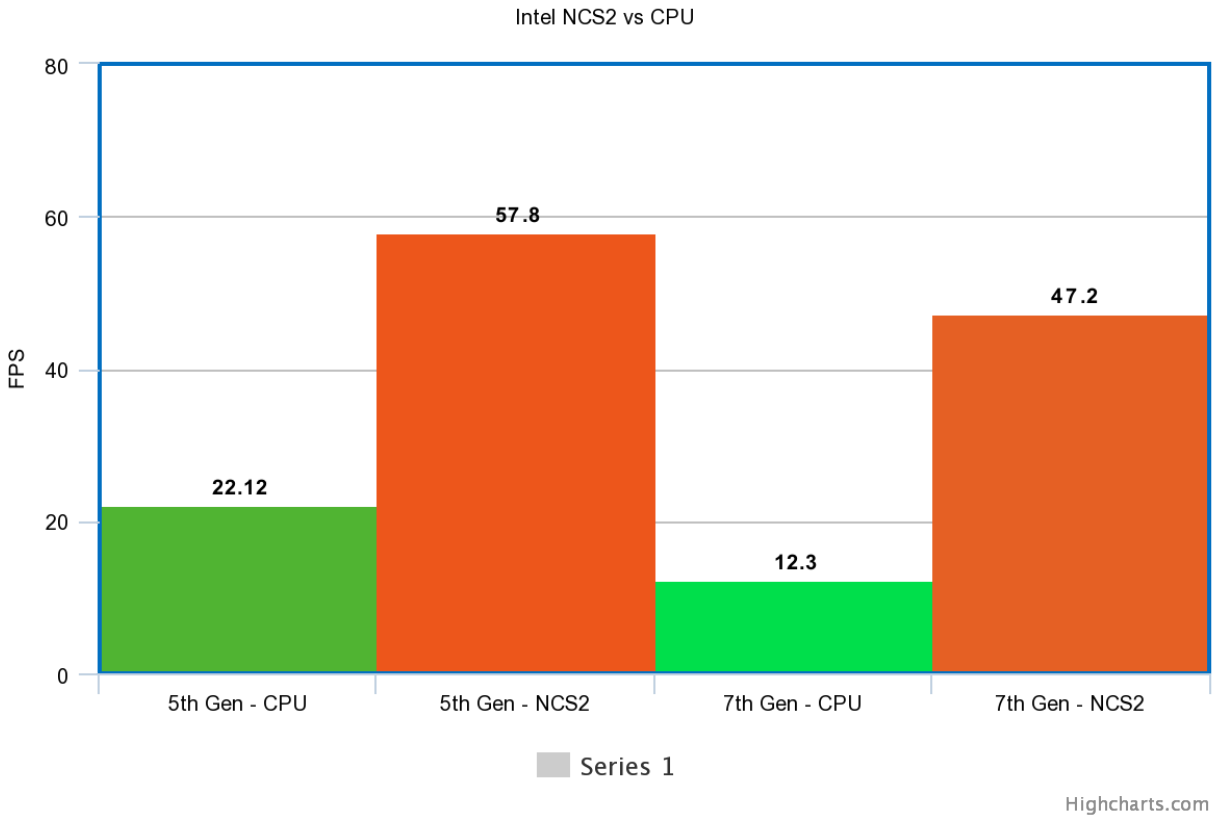


Performance testing

Inference is the action of applying a trained neural network to an input to generate an output. Inferencing can be done on many devices such as CPU and GPU. So, NCS2 would have to outperform the CPU of the platform it's plugged into to prove that it's usefull.

This testing is performed on both NCS2 and CPUs. GoogLeNet model for object dectection was run on both the Intel technologies. The model which was The CPUs it was compared to were:

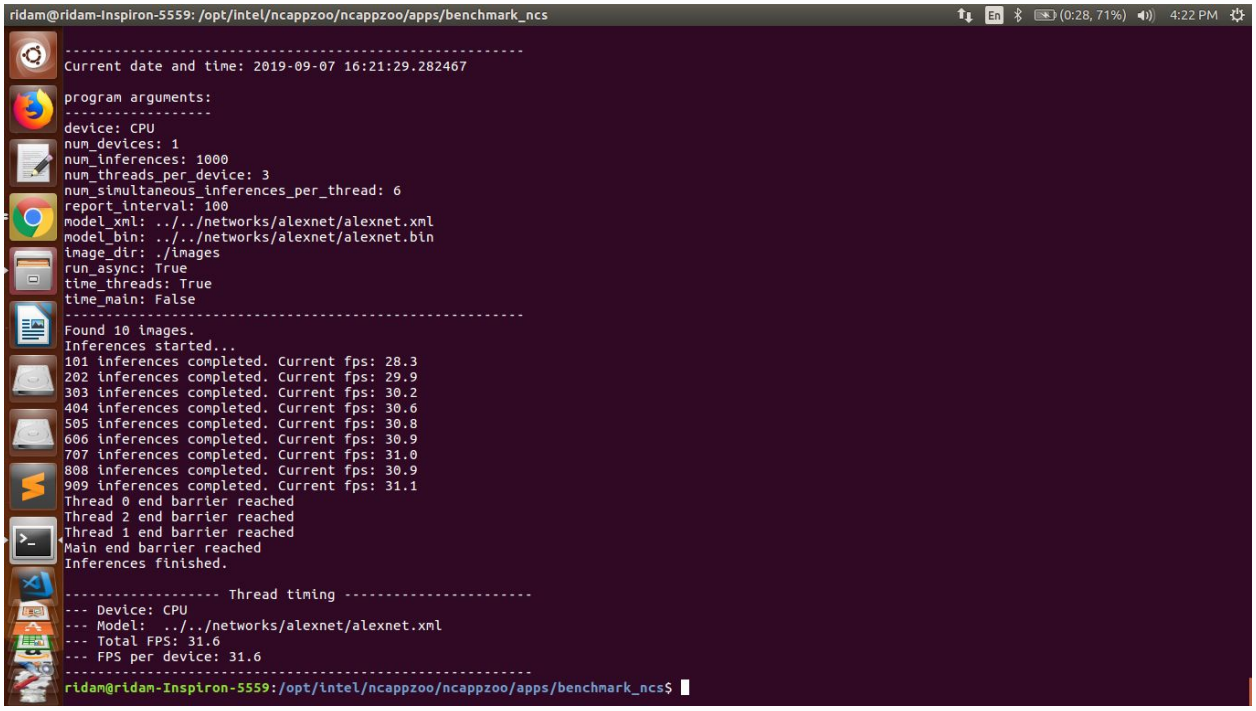
- Dell Inspiron 15 5000 series - 5th Gen Intel Core i5-5200U
- Dell Inspiron 5370 - Intel Core i5-8250U processor, 3.4 GHz base processor speed, 6MB cache



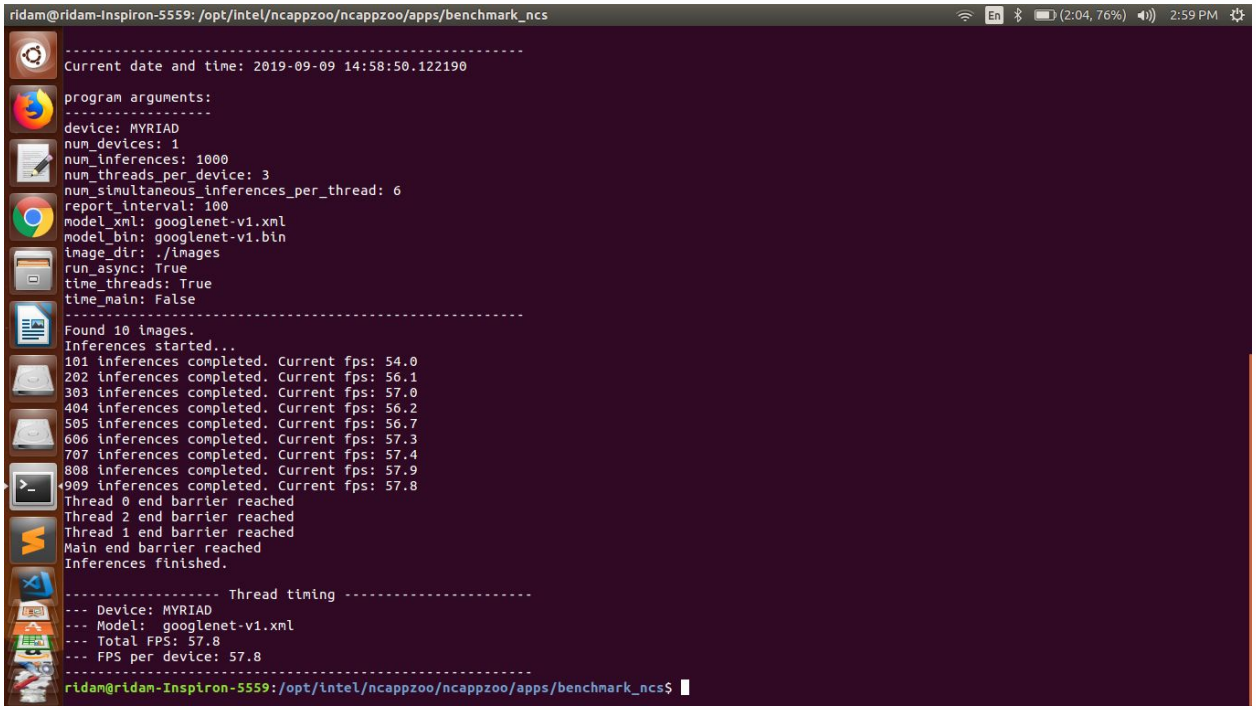
Model	CPU	NCS2
5th Gen	22.12 FPS	57.8 FPS
7th Gen	12.3 FPS	45.2 FPS

NOTE: We ran the above tests for a few times. Inorder to obtain more accurate results the above tests have to be performed multiple times.

NOTE: The floating point precision differs when running the tests on CPU and the NCS2. This is because the NCS2 only support half precision (FP16) whereas the CPU only support full, or normal, precision (FP32). But we are hoping that it wont make huge difference to inferencing.



OUTPUT - CPU(5th gen)



OUTPUT - NCS2 (5th gen)

Analysis

- The NCS2 falls behind when compared to Gen 5 and Gen 7 Intel i5 CPU.
- There are slight differences in results when the NCS2 is running on less powerful platforms vs. newer machines. This shows that the factors such as CPU, storage, memory also play a role in inference rather than just NCS2 itself.
- From the results it's clear that the Movidius NCS2 can't compete with a modern i7 or i5 CPU.
- But when compared to these CPUs, NCS2s are cheaper and draws less power.
- Many edge devices lack the capabilities to host a GPU, either due to space, cost or thermal limitations in such cases these NCSs can be used.
- The VPU solutions are intended for edge applications where compute power is limited and low-power is required. Example applications include drones, surveillance camera and virtual reality (VR) headsets.
- The most important thing to remember is that the NCS2 will be used to run Machine Learning frameworks like Caffe and Tensorflow and to leverage CNN (Convolutional Neural Networks) to do inferencing on data but not training the models on the data.

- We would still need GPUs or TPUs to train high end models.

Conclusion

The Intel Movidius Neural Compute Stick 2 is the second generation of deep learning development kit from Intel that is claimed to provide 8 times performance gain compared to its predecessor Modivus NCS. It acts as a harware accelerator for running inferencing at the edge. It is not as powerful as a full-on GPU nor a modern CPU. But it has potential to excel with edge applications where compute power is limited and low-power is required where the onboard CPU isn't powerful enough to do inferencing on its own.