HCL

Optimized Edge Analytics using Intel® Distribution of OpenVINO™ toolkit

WHITEPAPER

TABLE OF CONTENTS

Executive Summary

As the backbone technology of machine learning, deep neural networks (DNNs) have demonstrated their utility in a wide range of applications and problem sets. Running DNNs on resource-constrained edge devices is, however, by no means trivial, since it incurs high performance and energy overhead. Offloading DNNs to the cloud for execution may result in unpredictable performance, due to the uncontrolled long wide-area network latency. This paper describes a use case of optimizing a DNN model deployed on Intel-powered edge devices. Inferencing native DNN models on a low-powered edge device can be time consuming. Thus, by using the Intel® Distribution of OpenVINO™ toolkit, the DNN model was optimized to decrease the inference time. In this use case, an Inception v2 based SSD object detection model was optimized for inference which gave 3X performance boost when tested on an Intel Xeon Processor-based platform.

Introduction

Edge analytics is the process of collecting processing, and analyzing of data at the edge of a system either at or near a sensor, a network switch or some other connected device. With the advancement of Internet of Things (IoT), numerous industries, for example, retail, manufacturing, transportation, and energy are generating vast amounts of data at the edge of the network. Edge analytics offers few key benefits:

- Reduces latency issues in inferencing
- Enables organizations to scale their analytics processing capabilities
- Reduces overall expenses by minimizing bandwidth, scaling of the operations and reducing the latency of critical decisions

The most important part lies in the deployment after training a DNN model into a production environment. Sometimes, the production environment is not at par with the training environment which generates a challenge in terms of designing a process pipeline which utilizes the prediction from the model. The time taken to run a prediction on the model affects real-time execution of the code, challenging the idea of using a Deep model with higher accuracy. This forces to use specialized computing hardware designed specifically to parallelize the compute load and boost the inference time. More often these compute devices are costly and power-hungry. Thus, the solution is based on Intel® Distribution of OpenVINO™ toolkit to optimize the DNN model specifically for the underlying hardware to shorten the inference time.

Object Detection on Edge

Traditional methods for deploying Neural Network into a production environment involves multithreading concept, which is designing a specific load balancer, using a smaller model and architectural level changes. In the threading and load balancer-based pipeline the capabilities of the underlying hardware are used to parallelize the process execution. It involves having separate threads to load the data and run prediction to execute the control loop and utilize the prediction to make a decision. While the other two ideas involve modifying the Net in terms of computational complications. The above-stated methods require a sophisticated understanding of the environment in which the deployment will be done, Neural net model and the data source. This can get complicated when finding the ideal configuration that works for the specific task. Table 1 shows the comparison on different hardware before and after optimization with Intel® Distribution of OpenVINO™ toolkit.

	Before Intel® Distribution of OpenVINO™ Toolkit				After Intel® Distribution of OpenVINO™ Toolkit	
Hardware	Intel® Xeon® Platinum 8153 CPU @ 2.00GHz	Intel® Xeon® Platinum 8256 CPU @ 3.80GHz	Intel® Xeon® Platinum 8153 CPU @ 2.00GHz		Intel® Xeon® Platinum 8256 CPU @ 3.80GHz	
			FP32	INT8	FP32	INT8
Inference time (in mms)	1X	1X	4X	4X	3X	4X
Compute FPS	1X	1X	4X	4X	3X	4X

Table 1. Inference Time Before and After Intel® Distribution of OpenVINO™ toolkit

Object detection is one such problem where Neural Nets are excelling when compared to the traditional computer vision-based approach with improved accuracy. DL based detection models are performing well on the high-end computation device. It poses a really interesting benefit when deployed at devices like camera, security system, product scanner, etc. Even with the development in real time, single shot object detection NN techniques, the seamless integration of these methods with a low powered device is still tricky. Table 2 shows the comparison between different hardware on Intel IOT edge cloud.

Solution Approach

The solution optimizes an Inception v2 based SSD object detection model using Intel® Distribution of OpenVINO™ toolkit for inference on various edge devices. The Intel® Optimization for TensorFlow based model is converted to an Intermediate Representation (IR) model which can be used by Intel inference engine. The Intel inference Engine facilitates speeding up the execution time by selectively executing different layers on specific computational hardware available. Workload optimization, asynchronous execution, utilize custom layers on specialized hardware's are some of the optimization steps used in our work. Figure 1 shows the pipeline for Intel® Distribution of OpenVINO™ toolkit conversion.

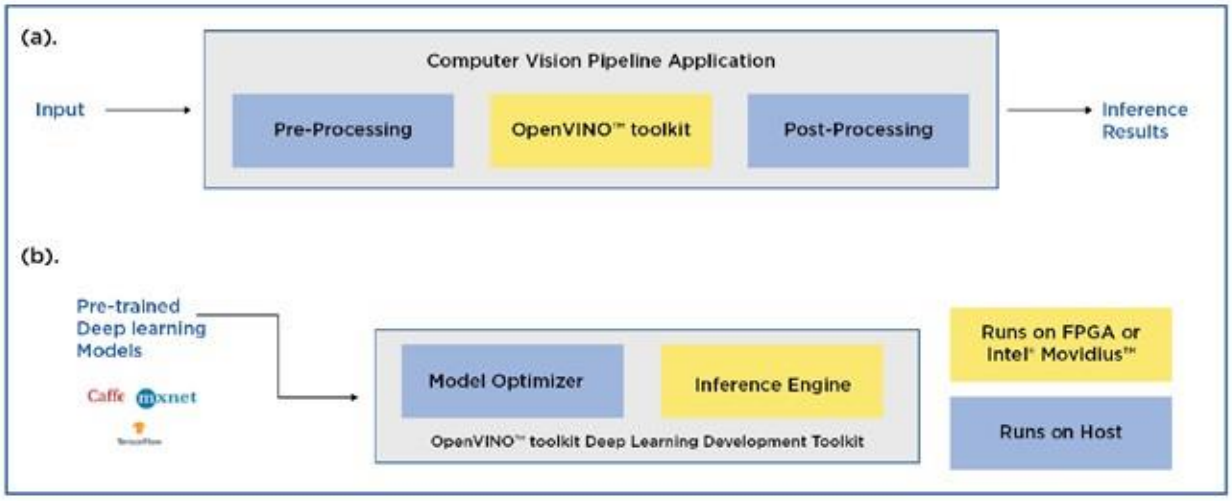


Fig 1. Intel® Distribution of OpenVINO™ toolkit workflow for optimization and deploying a trained deep learning model

Results and Discussions

Table 1 tells us that Intel® Xeon® Platinum 8256 CPU @ 3.80GHz has the highest FPS value compared to Intel® Xeon® Platinum 8153 CPU @ 2.00GHz because it supports VNNI (Vector Neural Network instructions) that accelerate the convolution operation. Table 2 shows the comparison in improvement of inference time and model performance after optimization through Intel® Distribution of OpenVINO™ toolkit in different hardware.

Intel® Core™ i5-6500TE is 14nm in size with 6M cache memory which includes 4 cores and 4 threads with processor base frequency of 2.30 GHz and Max turbo frequency of 3.30 GHz. It supports 4K with a maximum resolution (HDMI 1.4) of 4096x2304@24Hz, DP of 4096x2304@60Hz and eDP – integrated flat panel of 4096x2304@60Hz.

Intel® Xeon® Processor E3-1268L v5 is also a 14nm in size but with 8M cache memory which has 4 cores and 8 threads with processor base frequency of 2.40 GHz and Max turbo frequency of 3.40 GHz. It supports 4K at 60Hz with a maximum resolution (HDMI 1.4) of 4096x2304@24Hz, DP of 4096x2304@60Hz and eDP – integrated flat panel of 4096x2304@60Hz.

Intel® Core™ i5-6500TE CPU with IEI Mustang-F100-A10 FPGA is powered by Open Visual Inference & Neural Network Optimization (OpenVINO) toolkit. It supports Ubuntu 16.04.3 LTS 64bit, CentOS 7.4 64bit, Windows 10. Its high flexibility allows trained data such as Caffe, Intel® Optimization for TensorFlow, and MXNet to execute on it after converting to optimized IR. It has a memory of 8G on board DDR4 which consumes power of 40W.

Intel® Core™ i5-6500TE CP with Intel neural stick2 develops, fine tune and deploy convolutional neural networks (CNNs) on low-power applications that require real-time inferencing. It supports Raspberry Pi and heterogeneous execution across computer vision accelerators—CPU, GPU, VPU, and FPGA—using a common API. The operating systems that is supports include Ubuntu 16.04.3 LTS (64 bit), CentOS* 7.4 (64 bit), Windows 10 (64 bit), Raspbian (target only) and Other (via the open source distribution of Intel® Distribution of OpenVINO™ toolkit).

Intel® Core™ i5-6500te CPU with IEI Mustang-V100-MX8 is also powered by (Intel® Distribution of OpenVINO™ toolkit) toolkit. The operating system that it supports include Ubuntu 16.04.3 LTS 64bit, CentOS 7.4 64bit and Windows 10. It consumes approximately 25W or less power and has dataplane interface PCI Express x4 and is compliant with PCI Express Specification V2.0.

Intel® Atom® x7-E3950 Processor is also a 14nm in size but with 2M cache memory which has 4 cores and 4 threads with processor base frequency of 1.60 GHz and burst frequency of 2.0 GHz. It supports 4K at 60Hz with a maximum resolution (HDMI 1.4) of 3840x2160@30Hz, DP of 4096x2160@60Hz and eDP – integrated flat panel of 3840x2160@60Hz. It has 8 USB ports of 2.0/3.0.

Conclusion

The objective of this work was to enable DNN-based deep learning inferencing on edge devices with Intel CPUs, for optimization in real time. It is important to deploy an optimized version of a Neural Network in the production environment as it improves the throughput and latency of the process pipeline. In this case, we have used an object detection model based out of SSD inception v2 where we achieved 3X improved inference time while retaining the mAP value on Intel® Xeon® Processors using Intel® Distribution of OpenVINO™ toolkit.

References

• https://software.intel.com/en-us/openvino-toolkit/deep-learning-cv

• https://ark.intel.com/products/88186/Intel-Core-i5-6500TE-Processor-6M-Cache-up- to-3-30-GHz-

• https://ark.intel.com/products/88178/Intel-Xeon-Processor-E3-1268L-v5-8M-Cache-2-40-GHz-

• https://www.ieiworld.com/mustang-f100/en/

• https://software.intel.com/en-us/neural-compute-stick

• https://www.ieiworld.com/mustang-v100/en/

• https://ark.intel.com/products/96488/Intel-Atom-x7-E3950-Processor-2M-Cache-up- to-2-00-GHz-

• https://www.kdnuggets.com/2017/10/edge-analytics.html

Author Information

Guda Ramachandra.K.S:
Guda Ramachandra currently works as Data Scientist in Artificial Intelligence with deep expertise in Deep learning/Machine learning. He is an SME in developing Deep Learning computer vision Solutions. Guda Ramachandra has published 12 white papers in international forums/conferences.

Vijaya Yuvaram Singh.V.M:
Vijaya Yuvaram Singh works as Lead Engineer at Analytics COE HCL. He is an expert in Artificial Intelligence with Deep Learning ,Machine Learning and Robotics background.

ABOUT HCL TECHNOLOGIES

HCL Technologies (HCL) is a leading global IT services company that helps global enterprises re–imagine and transform their businesses through digital technology transformation. HCL operates out of 32 countries and has consolidated revenues of US$ 6.97 billion, for 12 months ended 31st December, 2016. HCL focuses on providing an integrated portfolio of services underlined by its Mode
1–2–3 growth strategy. Mode 1 encompasses the core services in the areas of Applications, Infrastructure, BPO, and Engineering & R&D services, leveraging DRYiCE™Autonomics to transform clients' business and IT landscape, making them 'lean' and 'agile'. Mode 2 focuses on experience–centric and outcome– oriented, services such as Digital and Analytics Services (BEYONDigital™), IoT WorKS™, Cloud and Security, utilizing DRYiCE™ Orchestration to drive business outcomes and enable enterprise digitalization. Mode 3 strategy is ecosystem–driven, creating innovative

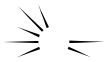
IP–partnerships to build products and platforms business.

HCL leverages its global network of integrated co-innovation labs, and global delivery capabilities to provide holistic multi–service delivery in key industry verticals including Financial Services, Manufacturing, Telecommunications, Media, Publishing, Entertainment, Retail CPG, Life Sciences Healthcare, Oil & Gas, Energy & Utilities, Travel, Transportation & Logistics and Government. With
120,000 professionals from diverse nationalities, HCL focuses on creating real value for customers by taking 'Relationships Beyond the Contract'. For more information, please visit www.hcltech.com.

ABOUT HCL ENTERPRISE

HCL is a $8.5 billion leading global technology and IT enterprise comprising two companies listed in India – HCL Technologies and HCL Infosystems. Founded in 1976, HCL is one of India's original IT garage start-ups. A pioneer of modern computing, HCL is a global transformational enterprise today. Its range of offerings includes product engineering, custom & package applications, BPO, IT infrastructure services, IT hardware, systems integration, and distribution of information and communications technology (ICT) products across a wide range of focused industry verticals. The HCL team consists of over 137,000+ ideapreneurs of diverse nationalities, who operate from 44 countries including over

500 points of presence in India. HCL has partnerships with several leading global
1000 firms, including leading IT and technology firms. For more information, please www.hcl.com



Hello there! I am an Ideapreneur. I believe that sustainable business outcomes are driven by relationships nurtured through values like trust, transparency and flexibility. I respect the contract, but believe in going beyond through collaboration, applied innovation and new generation partnership models that put your interest above everything else. Right now 137,000+ Ideapreneurs are in a Relationship Beyond the Contract™ with 500 customers in 44 countries. How can I help you?

HCL