WHITE PAPER

Healthcare
Intel® Xeon® Processors
Intel® Distribution of OpenVINO™ Toolkit

GE Healthcare's AIRx™ Tool Accelerates Magnetic Resonance Imaging using Intel® AI Technologies

Software optimizations, including the Intel® Distribution of OpenVINO™ Toolkit, produce a 4.3x speedup for identifying and aligning MR scans for diagnostic neuroimaging on Intel® Xeon® Processor-based systems.[1]

Authors

Aruna Narayanan
GE Healthcare

Mallory Busso
GE Healthcare

Zac Slavens
GE Healthcare

Valentina Taviani
GE Healthcare

Dmitry Rizshkov
Intel Corporation

G. Anthony Reina
Intel Corporation

Prashant Shah
Intel Corporation

Table of Contents

Introduction

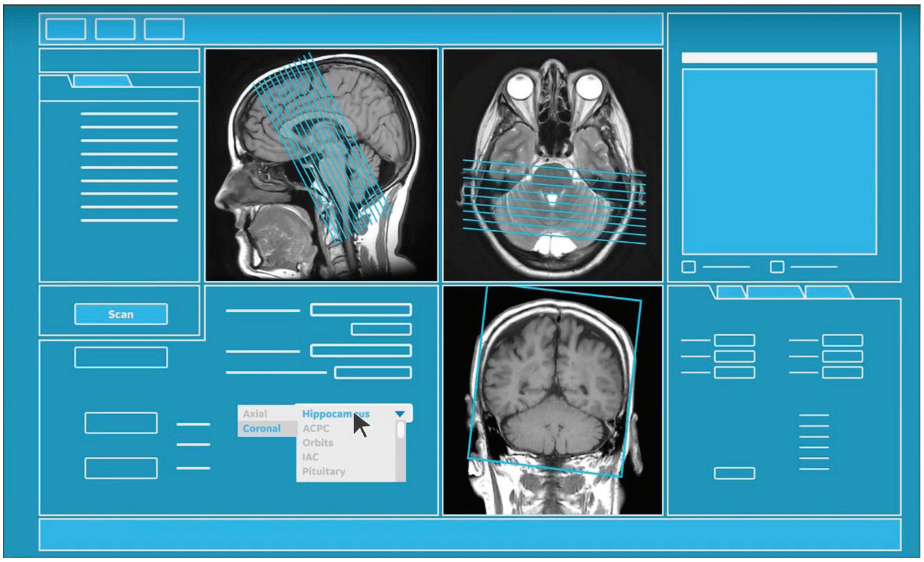


Figure 1. GE Healthcare's AIRx™ automatically aligns MR scans for better diagnostic imaging.

GE Healthcare's Artificial Intelligence Prescription (AIRx™) is an automated workflow tool for magnetic resonance (MR) brain scanning that has received 510(k) clearance from the FDA. Built on GE Healthcare's Edison platform, AIRx uses state-of-the-art artificial intelligence (AI) to precisely identify and align MR scans for diagnostic neuroimaging (Figure 1). AIRx automates slice prescriptions and reduces redundant, manual steps by using AI algorithms built into the MR technologist's existing workflow. This enhanced process allows for consistent, repeatable scan alignment to help physicians better monitor a patient across longitudinal studies which may be several months apart. GE Healthcare estimates that using AIRx may reduce the set-up time for an MR study by 40 to 60% while increasing accuracy and consistency of the scans.

GE Healthcare partnered with Intel to optimize the inference speed of AIRx on their existing Intel® Xeon® CPU-based platforms. Using software optimizations, including the Intel® Distribution of OpenVINO™ Toolkit, GE Healthcare was able to reduce the total inference time from 2.85 seconds down to 0.659 seconds without the additional cost of accelerators— improving patient care without increasing healthcare costs[1].

Topology and Data

AIRx relies on 11 separate convolutional neural network (CNN) topologies which consider both 3-dimensional and 3-plane (axial, coronal, sagittal) 2-dimensional MR scans. CNNs are a class of deep learning networks commonly applied to analyzing visual data. These 2D and 3D Cascaded CNN models are combined to automatically detect the scanning target (e.g. hippocampus, pituitary, circle of Willis) and determine the best alignment for the scanning protocol.

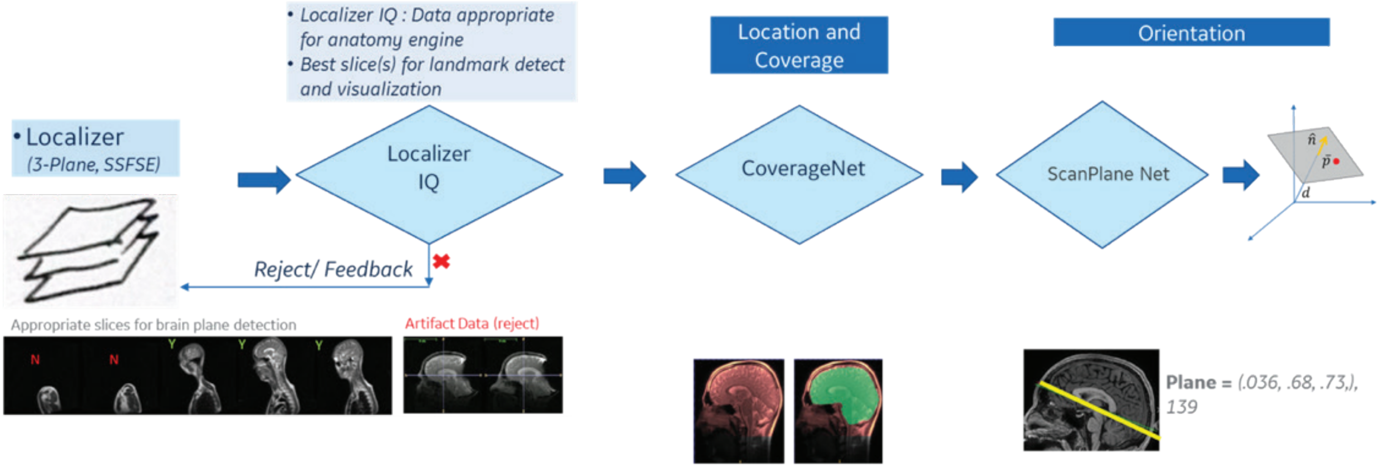

• Localizer IQ : Data appropriate for anatomy engine
• Best slice(s) for landmark detect and visualization
Location and Coverage
Orientation
• Localizer (3-Plane, SSFSE)
Localizer IQ
CoverageNet
ScanPlane Net
$\hat{n}$
$\vec{p}$
d
Reject/ Feedback
Appropriate slices for brain plane detection
N
N
Y
Y
Y
Artifact Data (reject)
Plane = (.036, .68, .73,), 139

Figure 2. AIRx™ workflow pipeline. DL-based intelligent scan placement framework. LocalizerIQ-Net classifies if the scan is suitable for the target anatomy. Coverage-Net determines the extent of the target anatomy in the scan, and Orientation-Net determines the best orientation and location of the target anatomy.

The total workload was divided into three distinct tasks as shown in Figure 2:

1. **LocalizerIQ-Net** is a 5-layer, dyadic reduction 2D CNN which determines if a given localizer image (coronal, axial, sagittal) is suitable to identify the target anatomy. If LocalizerIQ-Net determines an unsuitable localizer scan, then it prompts the MR technologist to repeat the initial scout scan.

2. **Coverage-Net** is a semantic segmentation CNN (2D U-Net variant) which determines the extent of the target anatomy within the localizer scan. This makes the algorithm generalizable to variations in patient anatomy. It consists of 3 models executed in parallel that cover the coronal, axial, and sagittal planes of the MR.

3. **Orientation-Net** is a 3D U-Net CNN which determines the best orientation and location to target the desired anatomy from the localizer images. This consists of three sub-models: two mid-sagittal plane (MSPNet) models (axial and coronal), executed in parallel so we only consider the maximal latency for these two models, and one Anterior Commissure-Posterior Commissure (ACPC) model.

The proprietary GE Healthcare dataset was generated from a global study of more than 1,300 studies using both 1.5T and 3.0T GE SIGNA™ MR scanners. For 2D models accepting a 3D volume input (e.g. Coverage-Net) the input shape varied in size between 288x220 and 320x320 with anywhere between 9 to 20 slices of thicknesses from 1.0 mm to 4.0 mm each. For 3D models, the input shape was always 256x256 with 9 slices. Scans also varied in contrast and MR protocols. Expert radiologists and technicians created the ground truth labels. Figure 3 shows examples of the raw data and ground truth labels.

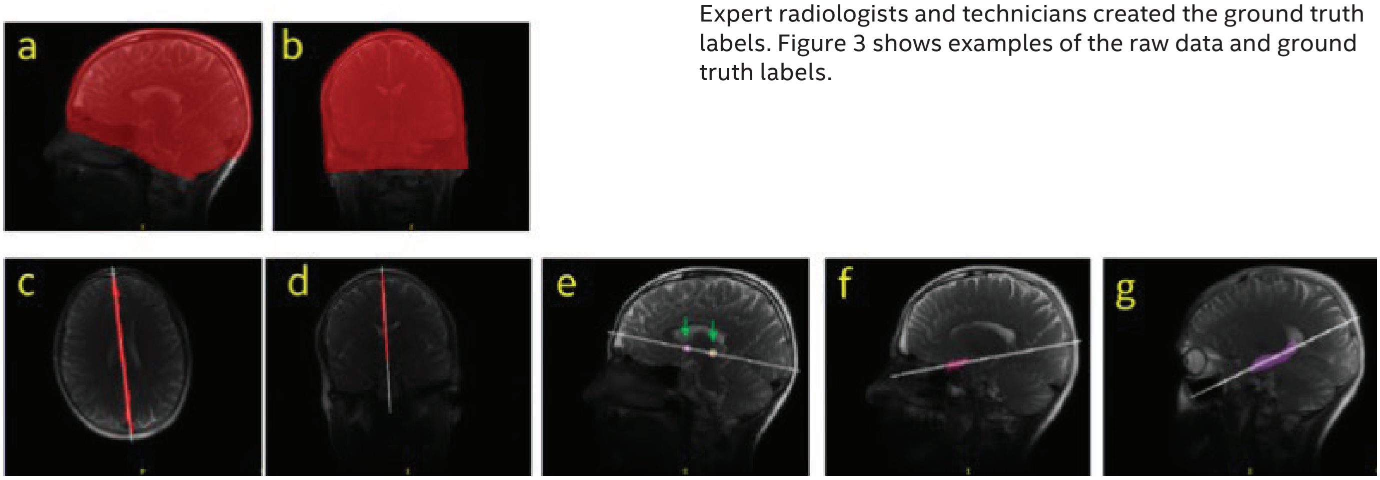


Figure 3. Examples of ground truth labels used for Coverage-Net and Orientation-Net. (a,b) Coverage-Net ground truth masks for sagittal and coronal localization scans. (c,d) Mid-sagittal plane (MSP) label (red) and the predicted plane (white) (e) Anterior commissure-Posterior commissure (ACPC) plane (f) optic nerve plane (g) hippocampal plane.

LocalizerIQ-Net was trained on 29,000 images (with online image augmentation) and tested on 700 images. It obtained a classification accuracy of 99.2%. Coverage-Net and Orientation-Net were trained on 21,770 3-D MR volumes (with online image augmentation) and tested on 505 volumes. Orientation-Net achieved a mean distance error of < 1 mm and a mean angle error of < 3° which was determined acceptable by expert radiologists. Prospective studies confirmed that the trained model produced excellent results when given new data (Figure 4), and was also robust for studies with significant pathology that otherwise would have been difficult to correctly orient.

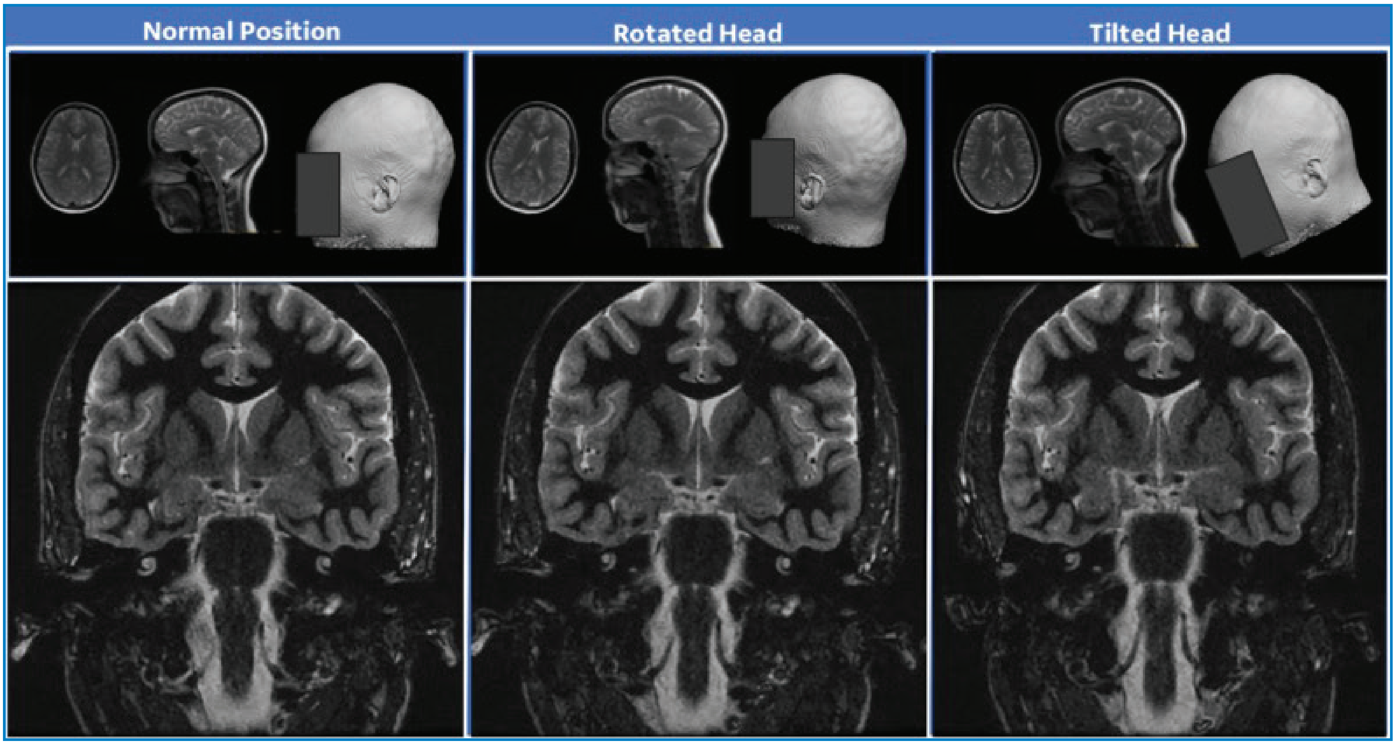


Figure 4. Prospective testing at new clinical sites showed that AIRx™ automatically produced consistent MR scans (bottom) regardless of patient positioning (top). This is desirable in longitudinal studies.

Benchmarking Metric

Several models within the workload could be run in parallel and take advantage of the multicore performance capabilities of Intel® CPUs. The remaining topologies were executed sequentially (Figure 2). The service level agreement for the workload was set at a maximum latency of 2.5 seconds for the total inference time of the Coverage-Net, MSPNet, and ACPC models on a GE Healthcare Gen6-P image compute node (3.10.0-862.el7.x86_64; Intel® Xeon® Processor E5-2680 v3, 2.5 GHz, 12 cores per socket, 2 sockets, 96 GB DDR4 RAM). GE Healthcare determined that this was the maximum allowable latency at which the MR technologist would not notice a disruption in routine workflow. The mean and standard deviation of 50 consecutive runs were recorded.

Optimizations

Intel® Distribution of OpenVINO™ Toolkit

Intel Distribution of OpenVINO Toolkit (2018 R4) was used to optimize the model for inference on Intel architectures. The toolkit's model optimizer is a Python* script that accepts a pre-trained TensorFlow* model (and other model formats), strips the TensorFlow-specific framework from the model, and performs several graph-level optimizations (Figure 5). The toolkit's Inference Engine leverages the multithreading operations from the Intel® Math Kernel Library for Deep Neural Networks (Intel® MKL-DNN) to take advantage of the Intel® Advanced Vector Extensions 512 (Intel® AVX-512) SIMD instructions found in modern multi-core Intel CPUs. Figure 6 illustrates demonstrate how layer fusion combines two or more operations within the same compute to reduce the number of separate compute operations during inference.

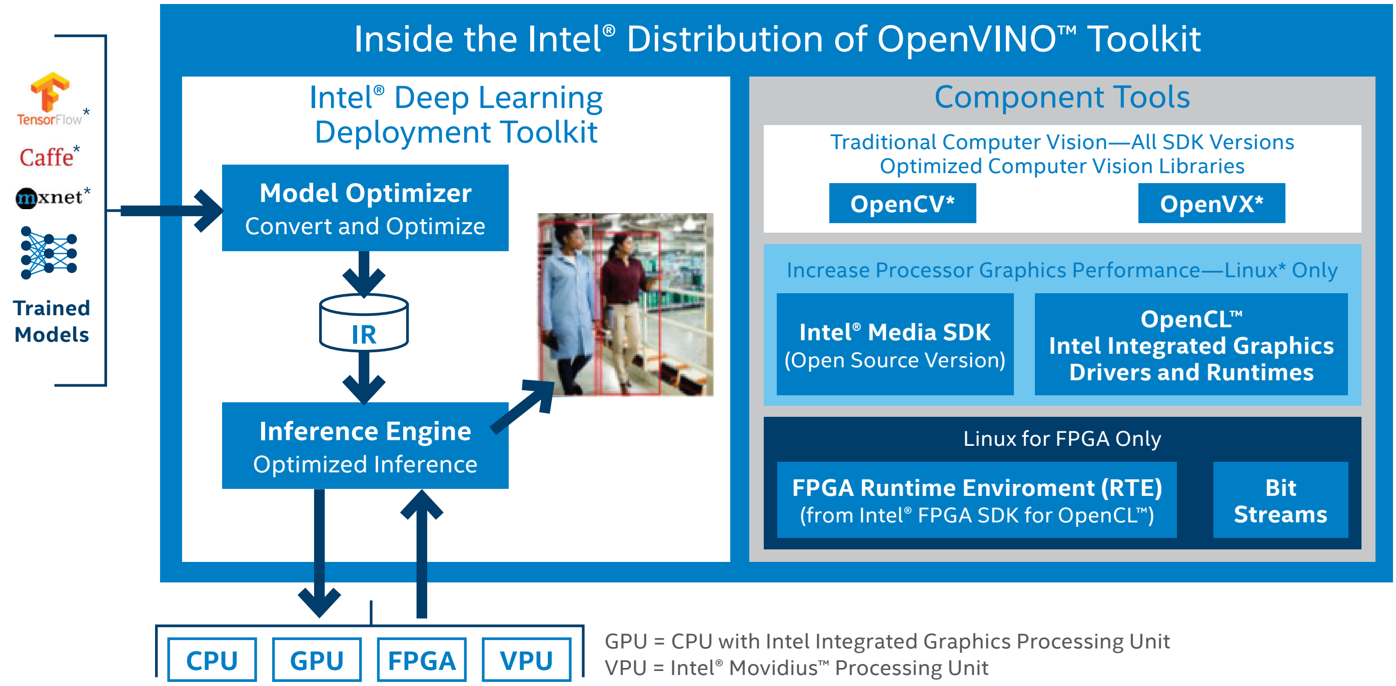

Inside the Intel® Distribution of OpenVINO™ Toolkit
TensorFlow*
Caffe*
mxnet*
Trained Models
Intel® Deep Learning Deployment Toolkit
Model Optimizer
Convert and Optimize
IR
Inference Engine
Optimized Inference
Component Tools
Traditional Computer Vision—All SDK Versions
Optimized Computer Vision Libraries
OpenCV*
OpenVX*
Increase Processor Graphics Performance—Linux* Only
Intel® Media SDK
(Open Source Version)
OpenCL™
Intel Integrated Graphics Drivers and Runtimes
Linux for FPGA Only
FPGA Runtime Enviroment (RTE)
(from Intel® FPGA SDK for OpenCL™)
Bit Streams
CPU
GPU
FPGA
VPU
GPU = CPU with Intel Integrated Graphics Processing Unit
VPU = Intel® Movidius™ Processing Unit

Figure 5. The Intel® Distribution of OpenVINO™ Toolkit's model optimizer transforms a pre-trained deep learning model for faster inference on Intel hardware.

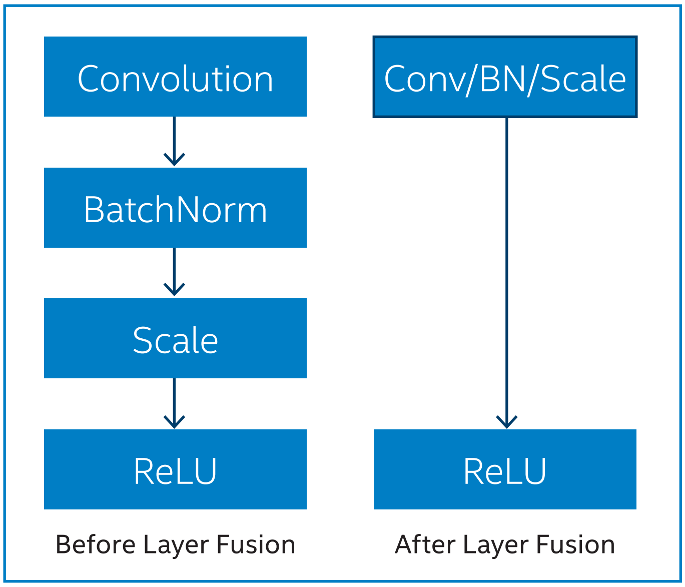
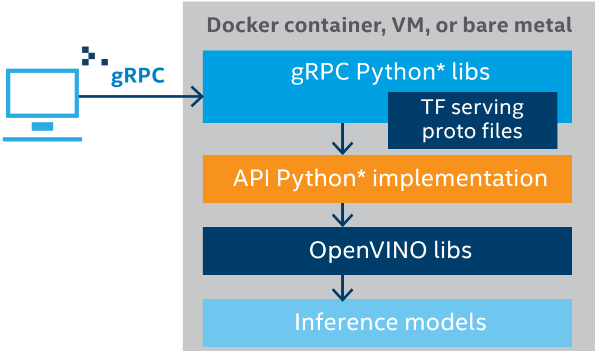
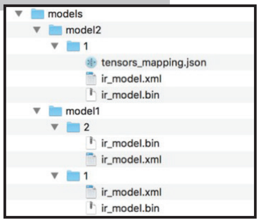


Figure 6. Layer Fusion. During inference operations such as batch normalization and scaling can be fused with the convolutional kernels and reduce the number of operations in the inference graph.

OpenVINO™ Model Server

GE Healthcare wanted to deploy multiple toolkit models into flexible, high performance, scalable components. To accomplish this, Intel developed the OpenVINO™ model server, a gRPC inference interface compatible with the TensorFlow Serving API that leverages the toolkit's faster inference engine for the backend. This speeds up the execution of the model on Intel CPUs and allows it to be used with other Intel® hardware, such as FPGAs and Intel® Movidius™ VPUs. The model server can be deployed on a bare metal server, a virtual machine, or a Docker container, making it suitable for a Kubernetes HPC environment where models can be served and load balanced across nodes.





```
▼ models
    ▼ model2
        ▼ 1
            tensors_mapping.json
            ir_model.xml
            ir_model.bin
    ▼ model1
        ▼ 2
            ir_model.bin
            ir_model.xml
        ▼ 1
            ir_model.xml
            ir_model.bin
```

Figure 7. OpenVINO Model Server

3D Deep Learning Primitives

A significant portion of the total AIRx workload relies on deep learning topologies that use 3D tensor operations such as 3D Convolution, MaxPooling, and ReLU. The additional tensor dimension greatly increases the memory footprint and computational complexity of these algorithms. The Intel Distribution of OpenVINO toolkit leverages Intel MKL-DNN's use of Intel AVX-512 single instruction multiple data (SIMD) hardware instructions which efficiently scales the operations across multiple CPU cores and balances data pre-fetching, cache blocking, and data formatting to promote optimal temporal and spatial locality of the data. Simply stated, these optimizations greatly improve the inference latency and throughput when compared with standard, non-optimized TensorFlow.

Results

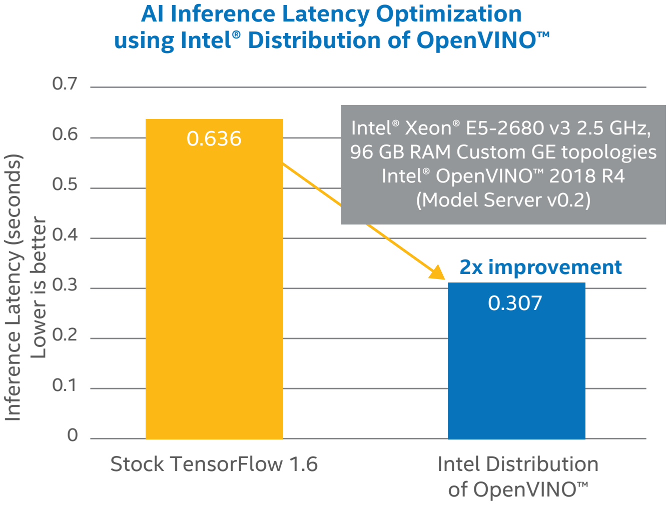


Figure 8. GE Healthcare achieved a 2x improvement in inference latency on the AIRx deep learning models by replacing the standard, unoptimized TensorFlow 1.6 models with the Intel OpenVINO models

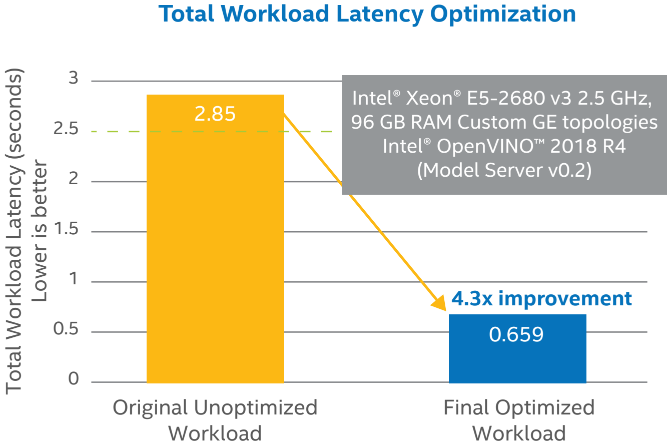


Figure 9. Total workload latency speed up achieved by overall software optimizations, including AI inference optimization shown in Figure 8

Figure 8 shows the improvement in the AI portion of the AIRx workload. By replacing the standard, unoptimized TensorFlow 1.6 models with the Intel OpenVINO models, the latency of the 2D and 3D deep learning models dropped from 0.636 seconds to 0.307 seconds. When combined with improvements in the non-AI portion of the workload—namely, optimizations in the Docker container structure, better multi-core/multi-threading balance, and changing the input of the model server from a pixel array to a binary format—the latency of the end-to-end AIRx workload dropped from 2.85 seconds down to 0.659 seconds. This is a 4.3x improvement in the end-to-end AIRx workload.

Conclusion

GE Healthcare's software optimizations, including the use of the Intel Distribution of OpenVINO™ toolkit, enabled them to optimize the inference speed of AIRx by over 4x on the Intel Xeon CPU architecture without the additional cost of accelerators, improving patient care without increasing healthcare costs. The OpenVINO Toolkit Model Server allows GE Healthcare to quickly scale the benefits of CPU inference across several product lines through their Edison platform. AIRx leverages the Edison platform to advance GE Healthcare's goal of offering intelligent applications and smart devices at the edge.

Notices and Disclaimers

Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors.

Performance tests, such as SYSmark and MobileMark, are measured using specific computer systems, components, software, operations and functions. Any change to any of those factors may cause the results to vary. You should consult other information and performance tests to assist you in fully evaluating your contemplated purchases, including the performance of that product when combined with other products. For more complete information visit www.intel.com/benchmarks.

Performance results are based on testing by GE Healthcare from November 2018 through January 2019 and may not reflect all publicly available security updates. No product or component can be absolutely secure. Intel does not control or audit third-party data. You should review this content, consult other sources, and confirm whether referenced data are accurate.

Optimization Notice: Intel's compilers may or may not optimize to the same degree for non-Intel microprocessors for optimizations that are not unique to Intel microprocessors. These optimizations include SSE2, SSE3, and SSSE3 instruction sets and other optimizations. Intel does not guarantee the availability, functionality, or effectiveness of any optimization on microprocessors not manufactured by Intel. Microprocessor-dependent optimizations in this product are intended for use with Intel microprocessors. Certain optimizations not specific to Intel microarchitecture are reserved for Intel microprocessors. Please refer to the applicable product User and Reference Guides for more information regarding the specific instruction sets covered by this notice.

Notice Revision #20110804

[1]Configurations:

Original model was trained using TensorFlow 1.6 for Python 2.7 without Intel optimizations and converted by GE Healthcare to OpenVINO 2018 R4.

HARDWARE USED FOR TESTING	
GE Gen6-P image compute node	3.10.0-862.el7.x86_64
Processor	Intel® Xeon® Processor E5-2680 v3
Speed	2.5 GHz
Cores	12 cores per socket, Docker container has access to 22 CPU cores
Sockets	2
RAM	96 GB (DDR4)
Hyperthreading	Enabled
Security Updates	Spectre and Meltdown Updates Applied

SOFTWARE USED FOR TESTING	
TensorFlow version	1.6 without Intel MKL-DNN optimizations
Gcc version	2.8.5
Python version	2.7
OpenVINO version	2018 R4 (Model Server v0.2)
OS	HeliOS 7.4 (Nitrogen)