SOLUTION BRIEF

Intel® AI Technology
Intel® Xeon® Scalable Processors

Improving Lung Cancer Detection with Advanced Intel® Technology

Intel® Distribution of OpenVINO™ Toolkit Accelerates Lung Nodule Detection and Segmentation on Intel® Xeon® Scalable Processor-based Platforms

Predible

"Predible Health's deep learning solutions have consistently demonstrated improved precision and efficiency for Radiologists, especially in cancer screening settings. Our collaboration with Intel enables us to deploy within the hospital premises, ensuring seamless workflow integration and real-time inference of the studies."

– Suthirth Vaidya, CEO, Predible Health

Authors

Predible Health
Kiran Vaidhya
Adarsh Raj
Krishna Chaitanya
Abhijith Chunduru
Suthirth Vaidya

Intel Corporation
Dr. Ramanathan Sethuraman
Madhu Kumar
Dmitry Rizshkov
Prashant Shah

Lung cancer, with an annual incidence of 3.34 million cases, is the deadliest form of cancer, with an estimated 1.88 million deaths per year worldwide[1]. Early detection is critical to long-term survival: stage 4 lung cancer has a 5-year survival rate of only 5%. But if caught at stage 1, patients experience a 5-year survival rate of 56%.

That's where technology comes in. The National Lung Screening Trial (NLST) revealed that participants who received low-dose helical CT (computed tomography) scans had a 20 percent lower risk of dying from lung cancer than participants who received standard chest X-rays.

Technology That Improves Patient Outcomes

Advances in multi-detector CT scanning have made high-resolution volumetric imaging possible in a single breath hold, at acceptable levels of radiation exposure. Several observational studies have shown that a low-dose helical CT scan of the lung detects more nodules and lung cancers, including early-stage cancers. Potentially malignant lung nodules can be identified from chest CT scans, and early intervention can result in a higher chance of long-term survival.

Solving the Detection Challenge

A typical chest CT scan contains between 300-500 slices, and a radiologist must examine each slice to detect lung nodules. Lung nodules are small masses of tissue in the lung that appear as round, white shadows on a CT scan; most are benign. They are often difficult to detect and document. Their detection requires specialized expertise, and, with widespread implementation of lung cancer screening programs, the burden on radiologists is rapidly increasing. Computer-aided-detection (CAD) is becoming increasingly useful in helping radiologists interpret high-dimensional imaging data like CT and MRI scans. CAD algorithms have also been successful in increasing radiologists' ability to detect lung nodules. With the advent of deep learning and convolutional neural networks (CNNs), CAD algorithms have started moving away from a reliance on hand-crafted features requiring custom engineering, to learning features from data through CNNs.

Predible and Intel Help Meet the Challenge

Predible Health and Intel Corporation have joined forces in the fight against lung cancer. Predible Health's deep learning algorithm for detecting lung nodules on CT scans has been optimized on powerful Intel® Xeon® Scalable processors using the Intel® Distribution of OpenVINO™ Toolkit.

Algorithm, Workflow, Requirements

Predible has built a software tool that automatically queries chest CT DICOM images from a Picture Archiving and Communication System (PACS), processes them, and uses neural networks to detect lung nodules. DICOM (Digital Imaging and Communications in Medicine) is an industry-standard method for handling, storing, printing, and transmitting medical imaging information. Once the DICOM series has been processed by the neural networks, the results are sent back to the PACS and are available to be viewed by the radiologist.

The deep learning system was trained to detect and segment lung nodules from chest CT scans. The system contains three stages, and each stage uses a combination of neural networks to solve a specific subtask.

Computational Requirements

For estimating computational requirements, Intel and Predible Health researchers assumed an operating window of one hour between start of image acquisition and start of interpretation by the radiologist. This means that the deep learning algorithms have about 30 minutes to process a chest CT scan and push the resulting secondary capture onto the PACS—which leaves 30 minutes for image acquisition. Hospitals may use either dedicated or shared compute assets for deep learning-based inferencing. In the former case, the expense of a dedicated compute asset could limit options for performance, resulting in a longer time to infer, while in the latter case, compute resource sharing could result in longer or shorter inference times, based on the clinical workflow.

For this study, the patches optimized by the Intel Distribution of OpenVINO toolkit for image recognition in workflow stages 1-3 are executed on Intel Xeon Scalable processors, and the performance is compared with a non-optimized PyTorch* software baseline on the same compute platform.

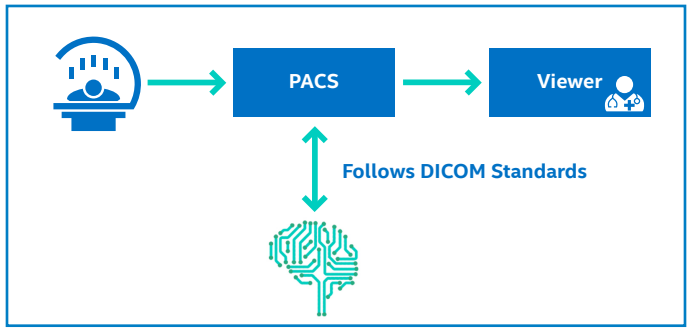

Figure 1. Automatic Query & Retrieval (AutoQR) pipeline integrates with PACS for processing Chest CT. Images used with permission by Predible Health.

Other pre-processing and post-processing stages are not compared since they are typically executed on Intel® processor-based systems that already deploy optimized libraries. Depending on the compute asset choice (dedicated versus shared), the pre- and post-processing stages can take several minutes. This leaves several minutes for the various deep learning stages to process.

Intel® Distribution of OpenVINO™ Toolkit Optimization

Enabled by tools like the Intel Distribution of OpenVINO toolkit, Intel Xeon Scalable processors offer a flexible platform for AI model inferencing.

The toolkit's offline Model Optimizer (MO) optimizes graph-level constructs such as node merging, batch normalization elimination and constant folding. The resulting output is an intermediate representation (IR) .xml file and a .bin file that contains the model weights. In the online process, the toolkit's Inference Engine optimizes MO output based on the target hardware: Intel® Xeon® processors, Intel® Core™

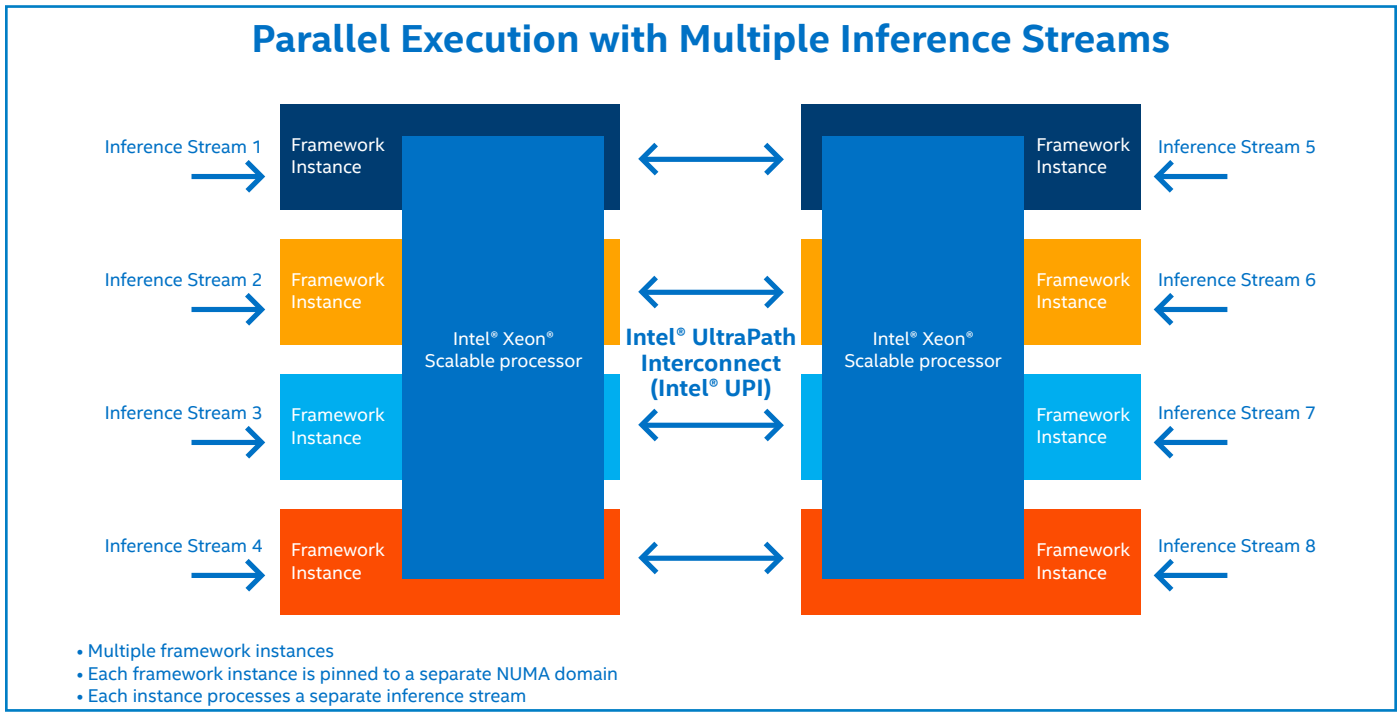

Figure 2. Sub-socket partitioning across dual-socket Intel® Xeon® processor-based platforms for multiple inference streams.

processors, Intel® Processor Graphics, an Intel® field-programmable gate array (FPGA), or Intel® Movidius™ Myriad™ vision processing units (VPUs). The Intel® Math Kernel Library for Deep Neural Networks (Intel® MKL-DNN) is an open-source performance library that significantly boosts performance of deep neural networks on Intel® CPUs.

Further performance gains can be obtained by running multiple instances of the toolkit on each of the sockets of a CPU (see Figure 2), instead of running just one instance of the toolkit in both sockets. Each instance is bound to one or more cores, which results in better core utilization. For example, as depicted above, eight instances of the toolkit are run on two Intel® Xeon® Scalable processor sockets.

Performance Comparison

Algorithmic Performance: The lobe segmentation model showed an average dice coefficient of 0.95 on Lung Tissue Research Consortium (LTRC) dataset. The nodule detection showed a performance of 89% sensitivity and specificity rate of one false positive per CT scan on a LIDC-IDRI dataset. The nodule segmentation model was trained and validated on a dataset from the Lung Image Database Consortium (LIDC) and Image Database Resource Initiative (IDRI). The model showed an average dice coefficient of 0.68 when compared against the intersection of contours annotated by the radiologists participating in the LIDC-IDRI study.

Compute Performance: Intel and Predible Health teams chose three different Intel Xeon Scalable processor SKUs as target implementation hardware. The Peak TFLOPS (single precision – FP32) for each processor are listed in Figure 3. All inferencing models use FP32 processing. Complete hardware and software configurations used for testing are provided in the Appendix.

Processor	Peak TFLOPS (FP32)
Intel® Xeon® Silver 4110 Processor	1.1
Intel® Xeon® Gold 6140 Processor	5.3
Intel® Xeon® Platinum 8168 Processor	8.3

Figure 3. Peak TFLOPS (FP32) for Intel Xeon CPUs.

As can be seen in Figure 4, the various modules in different stages of the Predible Lung CT model showed significant performance improvements.

Figure 5 depicts the number of layers per module in the Predible model before and after optimizations, clearly indicating the power of the toolkit.

Three major optimization steps contributed to the performance improvements:

1. Intel Distribution of OpenVINO toolkit-based inferencing model optimization
2. Multi-instance Intel Distribution of OpenVINO toolkit running on multiple CPU sockets
3. Custom inferencing model optimizations

Custom inferencing model optimizations provided opportunities to merge multiple layers or expose more channels for processing that can then benefit from better hardware utilization. These custom optimizations are enabled in subsequent releases, to benefit the broader AI model community.

Modules in Predible Lung CT model	Baseline (PyTorch)	Optimized (OpenVINO)
Stage 1 – Lobe Segmentation	62	46
Stage 2a.1 – Nodule Detection	138	69
Stage 2a.2 – Nodule Detection	138	69
Stage 2a.3 – Nodule Detection	111	61
Stage 2b.1 – False Positive Reduction	30	20
Stage 2b.2 – False Positive Reduction	32	20
Stage 3 – Nodule Segmentation	61	35

Figure 5. Number of layers per module before and after OpenVINO optimizations.

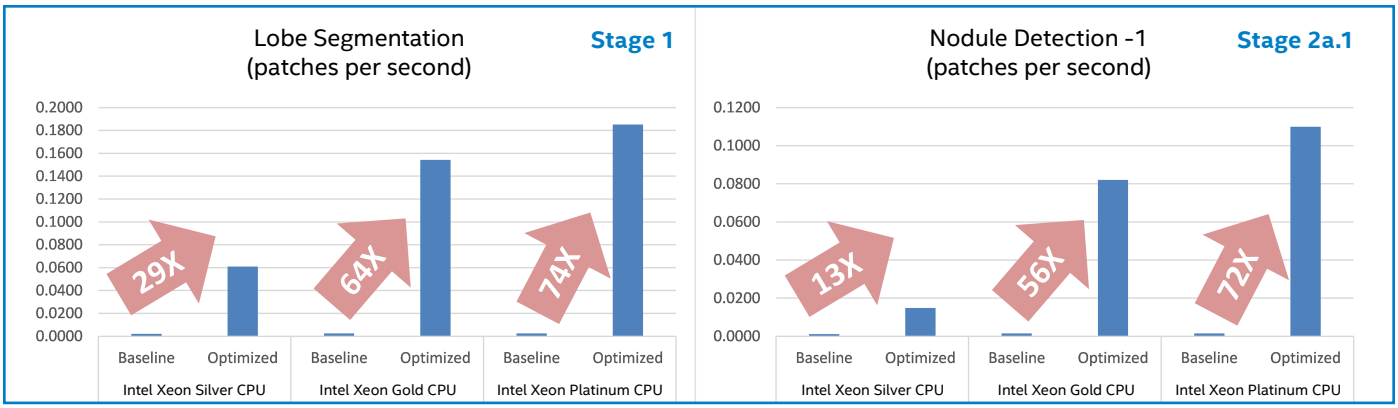

Figure 4. Performance of Intel Distribution of OpenVINO toolkit optimization versus PyTorch* baseline for Stages 1 and 2a.1 Similar improvements were found for other stages.

Table 1 describes the number of patches per stage of the Lung CT model and the overall time taken for the completion of each stage using the chosen Intel Xeon processors. Total time taken for all modules is under a minute for Intel Xeon Gold and Platinum CPUs and under 5 minutes for Intel Xeon Silver CPUs, clearly highlighting tradeoffs between latency and cost. For dedicated compute assets executing only one inferencing model, it may make sense to pick an entry processor that meets throughput needs while keeping costs low. For shared compute assets, where several inferencing models are executed concurrently and may have differing throughput and latency needs, Gold and Platinum parts offer the needed compute power and agility to handle concurrent invocations of models with differing performance needs. Also depicted in Table 1 is the choice of instance and batch size for each module in various stages of the Predible Lung CT model that provides the best performance on different CPUs.

Modules		Number of Patches (p)	**Intel® Xeon® Scalable Processor Family** (Intel® Xeon® Silver 4110 Processor / Intel® Xeon® Gold 6140 Processor / Intel® Xeon® Platinum 8168 Processor) [1.1 / 5.3 / 8.3] TFLOPS											
			Silver			Gold			Platinum			Overall time (secs.)		
Stage #	Function		Instance (i)	Batch size (b)	Time taken per (i*b) [sec]	Instance (i)	Batch size (b)	Time taken per (i*b) [sec]	Instance (i)	Batch size (b)	Time taken per (i*b) [sec]	Silver	Gold	Platinum
1	Lobe Segmentation	324	8	1	0.4	6	1	0.12	6	1	0.1	16.4	6.5	5.4
2a	Nodule Detection -1	84	4	1	3.2	6	1	0.87	12	1	1.3	67.2	12.2	9.1
	Nodule Detection -2	84	4	1	3.2	6	1	0.87	12	1	1.3	67.2	12.2	9.1
	Nodule Detection -3	84	4	1	2.41	6	1	0.64	4	1	0.31	50.6	9.0	6.5
2b	False Positive Reduction -1	100	4	1	1.05	6	1	0.29	6	1	0.2	26.3	4.9	3.4
	False Positive Reduction -2	100	4	1	1.05	6	1	0.28	6	1	0.2	26.3	4.8	3.4
3	Nodule Segmentation	30	2	1	2.22	6	1	1.19	2	1	0.3	33.3	6.0	4.5
Total time for all modules (secs.)												287.2	55.4	41.4

Table 1. Time taken per module in each stage of the Predible Lung CT model.

Conclusion

Intel and Predible Health researchers realized three main conclusions from their work:

- **Intel Distribution of OpenVINO Toolkit**-based optimizations yield significant speed ups (up to 83X) on **Intel Xeon Scalable processors** vs. baseline configurations.
- **Intel Xeon Scalable processors** offer a range of performance/price options to meet a variety of workload needs.
- Thanks to the power of **Intel AI technologies**, Predible Health's complex Lung CT model can be computed in less than a minute of processing time.

Learn More

- **Intel Xeon Scalable processors**
  https://www.intel.com/content/www/us/en/products/processors/xeon/scalable.html
- **Intel Distribution of OpenVINO Toolkit**
  https://software.intel.com/en-us/openvino-toolkit
- **Predible Health**
  http://prediblehealth.com/

Appendix A: Hardware and Software Test Configurations

	Platinum	Gold	Silver
Tested By	Intel	Intel	Intel
Test Date	02-07-2019	02-07-2019	02-07-2019
Platform	S2600STQ	S2600WFQ	S2600BPB
#Nodes	1	1	1
#Sockets	2	2	2
CPU	8168	6140	4110
Cores per socket/Threads per socket	24/24	18/18	8/8
Serial No cpu0	-	-	-
Serial No cpu1	-	-	-
ucode	0x200005e	0x200005e	0x200005e
HT	off	off	off
Turbo	off	off	off
BIOS Version	SE5C620.86B.00.01.0009.101920170742	SE5C620.86B.00.01.0009.101920170742	SE5C620.86B.00.01.0015.110720180833
System DDR Mem Config: Slots/Cap/Run-speed	12/16GB/2666	12/16GB/2666	12/16GB/2666
System DCPMM Config: Slots/Cap/Run-speed	-	-	-
Total Memory/Node	192GB	192GB	192GB
Storage-Boot	INTEL SSDSC2KB48 480GB	INTEL SSDSC2KB48 480GB	INTEL SSDSC2KB96 960GB
Storage-Application	-	-	-
NIC	-	-	-
PCH	-	-	-
Other HW (Accelerator)	-	-	-
OS	Ubuntu 16.04.6 LTS	Ubuntu 16.04.6 LTS	Ubuntu 16.04.6 LTS
Kernel	GNU/Linux 5.1.5-050105-generic x86_64	GNU/Linux 5.1.5-050105-generic x86_65	GNU/Linux 5.1.5-050105-generic x86_66
Mitigation Variants	Mitigated	Mitigated	Mitigated

Workload & Version	Stage 1	Stage 2a.1	Stage 2a.2	Stage 2a.3	Stage 2b.1	Stage 2b.2	Stage 3
Compiler	5.4.0	5.4.0	5.4.0	5.4.0	5.4.0	5.4.0	5.4.0
Libraries	MKL-DNN (OpenVINO™ inbuilt version)	MKL-DNN (OpenVINO™ inbuilt version)	MKL-DNN (OpenVINO™ inbuilt version)	MKL-DNN (OpenVINO™ inbuilt version)	MKL-DNN (OpenVINO™ inbuilt version)	MKL-DNN (OpenVINO™ inbuilt version)	MKL-DNN (OpenVINO™ inbuilt version)
Frameworks Version	OpenVINO™ 2019 R1	OpenVINO™ 2019 R1	OpenVINO™ 2019 R1	OpenVINO™ 2019 R1	OpenVINO™ 2019 R1	OpenVINO™ 2019 R1	OpenVINO™ 2019 R1
Dataset	PH dataset	PH dataset	PH dataset	PH dataset	PH dataset	PH dataset	PH dataset
Topology	3D Unet	3D Resnet	3D Resnet	3D Resnet	3D Wide Resnet	3D Wide Resnet	3D Unet
Batch Size	1	1	1	1	1	1	1

1 For references on this and other statistics, please see the *Deep Learning for Lung Cancer Detection* whitepaper by Intel Corporation and Predible Health, 2019.

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 intel Corporation as of July 10, 2019 and may not reflect all publicly available security updates. See configuration disclosure for details.

No product or component can be absolutely secure.

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/CVN/ACG