SOLUTION BRIEF

Health and Life Sciences

Siemens Healthineers Accelerates AI for Cardiology, Powered by 2nd Generation Intel® Xeon® Scalable Processors with Intel® Deep Learning Boost

Authors

Dr. Bogdan Georgescu, principal research scientist, Siemens Healthineers

Gianluca Paladini, senior director, Engineering, Siemens Healthineers

Prashant Shah, engineering director, Health and Life Sciences AI, Intel Corporation

Dmitry Rizshkov, engineer, Deep Learning Software, Intel Corporation

Christian Schmidt, global account executive, Intel Corporation

Dr. Dorin Comaniciu, senior VP, Artificial Intelligence and Digital Innovation, Siemens Healthineers

"Our ongoing collaboration with Intel has resulted in the acceleration of a variety of our deep learning models by several times on various modalities—from magnetic resonance to computed tomography—with current Intel® Xeon® Scalable processors. Furthermore, preliminary tests on 2nd Generation Intel Xeon Scalable processors with Intel® Deep Learning Boost technology and optimizations on our deep learning model with Intel® Distribution of OpenVINO™ toolkit have shown a 5.5x acceleration improvement from our baseline."[1]

—Dr. Dorin Comaniciu, senior VP, Siemens Healthineers

SIEMENS Healthineers

AI optimized with Intel® technologies gains 5.5x speedup for quantifying heart function in cardiac MRI[1]

Executive summary

The health and life sciences industry is digitizing healthcare and leveraging artificial intelligence (AI) to accelerate clinical workflows, help improve accuracy and diagnosis, reduce hospital costs, and support medical research. AI can quickly provide visibility into anatomical systems and identify abnormalities, which helps clinicians know where to focus patient care.

Radiologists and cardiologists are utilizing AI techniques such as object detection and segmentation to help identify and compare relevant patterns and other imaging data faster and more accurately. Taking advantage of AI applications requires new levels of system performance to keep up with the radiologist workflow. Siemens Healthineers and Intel are working to accelerate AI for cardiac imaging using 2nd Generation Intel® Xeon® Scalable processors with Intel® Deep Learning Boost and Intel® Distribution™ of OpenVINO™ toolkit.

Intel's leadership in AI hardware and software technology creates opportunities for a breadth of high-performance, highly accurate medical imaging applications and solutions that speed time to insight, bringing the benefits of AI to the health and life sciences industry.

Challenges

One-third of all deaths—34 deaths per minute and 18 million each year—are due to cardiovascular disease.[2] Cardiac MRI has established itself as a gold standard for evaluating heart function, heart chamber volumes, and myocardial tissue evaluation.[3] To extract quantitative measurements from the CMR images, the cardiologists typically use manual or semi-automatic tools, a time-consuming step that is error prone and affected by the inter-user subjectivity in interpreting the images. Adding to the challenge is the wide range of available cardiovascular data types coupled with physician subjectivity in interpreting and measuring the data, which can lead to undiagnosed or misdiagnosed diseases.

Siemens Healthineers is a pioneer in the use of AI for medical applications. They research novel AI use cases and incorporate their findings in their scanners and radiology and cardiology applications. These AI use cases need to be implemented seamlessly into the clinical workflow to save time and increase consistency and accuracy in measurements and diagnoses.

However, AI cannot come at the expense of delays in the clinical workflow. The computation system powering AI needs to keep pace with the data being generated by the scanners, requiring the system to offer low latency for AI inference and high throughput. This allows healthcare systems to care for more patients in a day.

Accelerators such as GPUs are often considered for AI workloads, but may add system and operational costs and complexity and prevent backward compatibility. Most systems deployed by Siemens Healthineers are already powered by Intel® CPUs; now Siemens Healthineers would like to leverage their existing CPU-based infrastructure to run AI inference workloads.

Siemens Healthineers is developing AI-based technologies for the analysis of cardiac magnetic resonance imaging (MRI) exams. One example is the automated delineation of cardiac chambers, which is foundational for many additional use cases, such as automated extraction of morphological and functional features from cardiac MRI for diagnosis and disease monitoring; quantitative functional analysis of the whole heart; and accurate quantification of the cardiac chamber volume, ejection fraction (EF), and myocardial mass.[1]

Siemens Healthineers was interested in accelerating the inference of the semantic segmentation model using 2nd Generation Intel Xeon Scalable processors with Intel Deep Learning Boost.

Solution

Siemens Healthineers and Intel have collaborated on optimizing Siemens Healthineers' heart chamber detection and quantification model for 2nd Generation Intel Xeon Scalable processors.[1] This AI model performs semantic segmentation of the left and right ventricles of the heart and can be extended to all four chambers. The input to the AI model is a stack of MRI images of the beating heart and the output identifies regions or structures of the heart, where each structure is color coded. This automates the labor-intensive manual segmentation process, accelerating time to results.

2nd Generation Intel Xeon Scalable processors offer an affordable, flexible platform for the inferencing of AI models, particularly in conjunction with tools like Intel Distribution of OpenVINO toolkit, which can help accelerate the development of high performance computer vision and deep learning inference into vision applications, without sacrificing the speed and accuracy of time-critical diagnosis and decision-making unique to healthcare.

> "We can now develop multiple near-real-time, often critical medical imaging use cases, such as cardiac MRI and others, using Intel® Xeon® Scalable processors, without the added cost or complexity of accelerators."
>
> —**Dr. Dorin Comaniciu**, senior VP, Siemens Healthineers

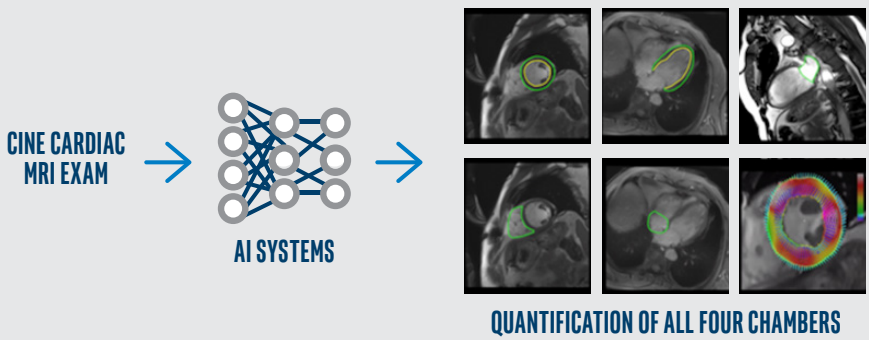
Intel Deep Learning Boost is built into 2nd Generation Intel Xeon Scalable processors to accelerate deep learning use cases. It extends the instruction set with a new Vector Neural Network Instruction (VNNI). Tasks such as convolutions, which typically required many instructions, can now be accomplished with just one instruction. Examples of these targeted workloads include image classification, image segmentation, speech recognition, language translation, object detection, and more. In order to take advantage of these instructions, models that are typically trained in floating point 32 (fp32) need to be quantized to int8. Through quantization, these workloads can be accelerated, but care must be taken to preserve the accuracy of the model.

The team used the Intel Distribution of OpenVINO toolkit to optimize, quantize, and execute the model. The resulting solution achieved 5.5x speedup with almost no degradation in accuracy.[1]

Such accelerations enable future solutions that:

- Process cardiac MRI data with unprecedented efficiency; at 200 fps, a full cardiac MRI exam, short axis spatio-temporal stack can be analyzed in less than a second
- Open the possibility for near-real-time clinical applications of cardiac MRI, making the interpretation of data available right after its acquisition

For a large subset of AI workloads, 2nd Generation Intel Xeon Scalable processors can better meet the product needs for incorporating AI models. This also enables medical equipment companies to offer AI solutions at lower cost to their customers. The high-performance platform capabilities eliminate the need for special accelerators.



SOLUTION COMPONENTS

- 2nd Generation Intel® Xeon® Scalable processors
- Intel® Deep Learning Boost
- Intel® Distribution of OpenVINO™ toolkit

Figure 1. Siemens Healthineers and Intel are accelerating AI analysis of cardiac MRI with 2nd Generation Intel® Xeon® Scalable processors, improving performance and optimizing AI workloads[1]

How it works

Achieving 5.5x speedup of int8 over the baseline fp32 model on 2nd Generation Intel Xeon Scalable processors is a result of efficient, low-precision convolutions due to Intel Deep Learning Boost, efficient concatenations in int8, and resample operation optimizations.[1]

The neural network is trained to identify regions of the heart. The weights and activations of the neural network are represented as fp numbers. The models are typically trained in fp32 precision. Once the desired level of accuracy is obtained, the model is ready to be incorporated in products. While many fp32 models are deployed in products today, quantizing the model to int8 can offer significant performance benefits. Typically, when quantization is performed correctly, there is little to no accuracy loss in the resulting model. Our goal was to limit the accuracy loss to < 0.5%. Our tests indicate that the resulting accuracy loss obtained was < 0.001%.

Intel quantized the trained model from Siemens Healthineers from fp32 to int8 using the Intel Distribution of OpenVINO toolkit, ensuring that accuracy was not compromised. The resulting shift in accuracy on the validation set of images was a low .001%; essentially maintaining imaging results for analysis at higher computational speeds.

Figure 2b shows a segmentation image for the heart. The AI model segments the various structures of heart. The ONNX output shows the solution running before quantification. The quantized int8 output displays nearly identical accuracy.

Figure 3 shows the overall process used to quantize the deep learning model for leveraging VNNI. Siemens Healthineers ONNX fp32 model for cardiac MRI segmentation is fed as input to the Model Optimizer, a component of Intel Distribution of OpenVINO toolkit. The resulting model has optimizations such as node merging, batch normalization elimination, and constant folding. This model is then used by the calibration tool along with a validation data set of images to generate an intermediate representation with statistics such as maximum and minimum activations per channel. This is needed for preserving accuracy during quantization. The process of generating internal representation is a one-time, offline process. At runtime, Intel Distribution of OpenVINO toolkit uses the statistics to execute the model in int8.

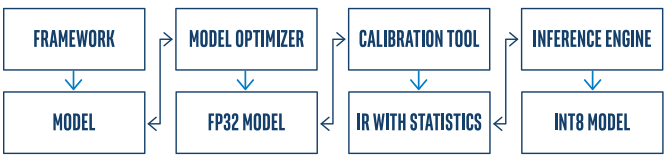


Figure 3. Quantization process with Intel® Distribution of OpenVINO™ toolkit

In addition to quantization, Intel and Siemens Healthineers conducted further optimizations. The team added int8 support to CPU extensions in Intel Distribution of OpenVINO toolkit, added the int8 data type to resample extension, and created a custom concat primitive for the Intel Distribution of OpenVINO toolkit CPU plugin to parallelize the operation.

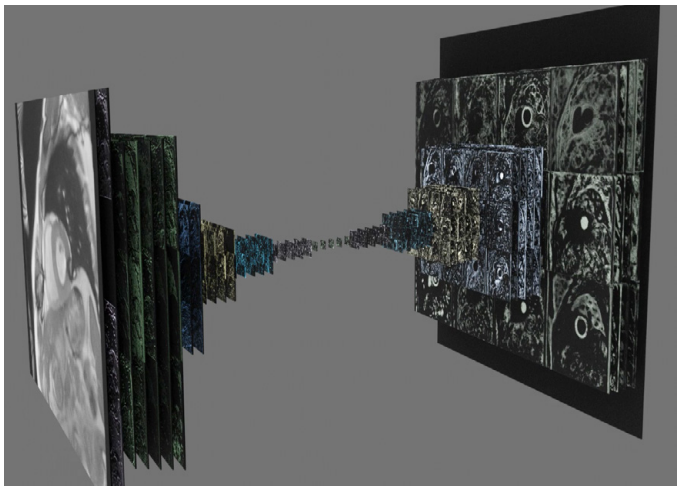


Figure 2a. Deep-Dense Net applied to Cine Cardiac MRI shows the hierarchical encoding of the image intensities to the decoding of the cardiac structures[4]

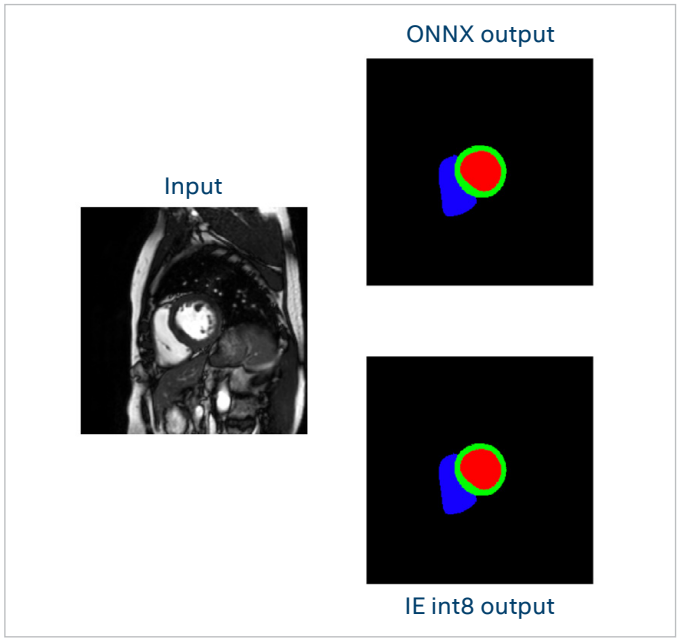


Figure 2b. Imaging results show negligible variation with improved performance

Results

With these optimizations, benchmarking the model on 2nd Generation Intel Xeon Scalable processors resulted in a 5.5x gain over the baseline model with overall throughput of 201.36 images/sec using 14 inference streams with two threads each, on a single 2nd Generation Intel Xeon Scalable processor socket.[1]

Dense-UNET	Batch Size	Throughput (Images/Sec) on 1S CLX 828		
		FP32	INT8	Speedup
Single stream	1	29.82	156.36	5.24x
Multistream (14 streams on a single 28-core CLX socket, with 2 threads per stream)	1	36.58	201.36	5.50x

These types of optimizations from Intel are enabling Siemens Healthineers to incorporate many near-real-time, critical medical imaging use cases on imaging solutions that already use Intel® processors to speed up time to solution without incurring the added cost or complexity of accelerators.

Conclusion

The optimization of the cardiac MRI segmentation model demonstrates the power of 2nd Generation Intel Xeon Scalable processors—allowing Siemens Healthineers to meet the growing needs of data-intensive AI applications for the health and life sciences industry. The optimization process indicates how solutions can be customized to meet specific real-world requirements for performance and accuracy.

Siemens Healthineers continues to refine and evolve AI training models to improve accuracy and support evolving workloads and use cases.

With Intel and Siemens Healthineers, the health and life sciences industry can leverage AI to integrate and analyze large amounts of data—helping to improve healthcare via more accurate diagnoses for better-targeted treatments.

Learn more

Explore 2nd Generation Intel Xeon Scalable processors ›

Visit the Resource and Design Center ›

Find out more about Intel Deep Learning Boost with VNNI ›

Download the Intel Distribution of OpenVINO toolkit ›

Discover Intel solutions for IoT ›

About Siemens Healthineers

At Siemens Healthineers, our purpose is to enable healthcare providers to increase value by empowering them on their journey toward expanding precision medicine, transforming care delivery, and improving patient experiences, all made possible by digitalizing healthcare.

An estimated 5 million patients globally benefit every day from our innovative technologies and services in the areas of diagnostic and therapeutic imaging, laboratory diagnostics, and molecular medicine, as well as digital health and enterprise services.

We are a leading medical technology company with over 120 years of experience and 18,000 patents globally. Through the dedication of more than 50,000 colleagues in 75 countries, we will continue to innovate and shape the future of healthcare.

Discover Siemens Healthineers' medical imaging solutions ›

Articles

Artificial Intelligence for Personalized Modeling of Cardiac Physiology, T. Mansi, D. Neumann, T. Passerini, D. Comaniciu, CBME 2017.

Automatic Delineation of Left and Right Ventricles in Cardiac MRI Sequences Using a Joint Ventricular Model, X. Lu, Y. Wang, B. Georgescu, A. Littman, D. Comaniciu, Functional Imaging and Modeling of the Heart (FIMH), New York, NY, 2011.

Towards Personalized Cardiology: Multi-Scale Modeling of the Failing Heart, E. Kayvanpour et al., PLOS ONE, 10(7), 2015.

Model Based Automated 4D Analysis for Real-Time Free-Breathing Cardiac MRI, B. Georgescu, T. Mansi, X. Lu, A. Kamen, D. Comaniciu, N. Seiberlich, M. Griswold, ISMRM, 2013.

Patient-Specific Modeling of the Whole Heart Anatomy, Dynamics and Hemodynamics from 4D Cardiac CT Images, V. Mihalef et al., Interface Focus Journal of the Royal Society, 2011.

1. This Siemens Healthineers' feature is currently under development and not available for sale. **5.5x speedup:** Based on Siemens Healthineers and Intel analysis on 2nd Gen Intel® Xeon® Platinum 8280 Processor (28 Cores) with 192GB, DDR4-2933, using Intel® OpenVino™ 2019 R1. HT ON, Turbo ON. CentOS Linux release 7.6.1810, kernel 4.19.5-1.el7.elrepo.x86_64.Custom topology and dataset (image resolution 288x288). Comparing FP32 vs Int8 with Intel® DL Boost performance on the system.
2. *Journal of the American College of Cardiology*, 2017.
3. *Recent Advances in Cardiovascular Magnetic Resonance Techniques and Applications*, Circ. Cardiovasc Imaging. 2017 Jun; 10(6).
4. *Rationale and Design for the Defibrillators to Reduce Risk by Magnetic Resonance Imaging Evaluation (DETERMINE) Trial*, A. H. et al., J Cardiovasc Electrophysiol, 20(9):982–7, 2009.

0319/MW/CMD/PDF 338825-001US