SOLUTION BRIEF

AI Model Development
Neural Networks

Huiying Medical Technology Optimizes Breast Cancer Early Screening and Diagnosis with Intel® AI Technologies

Intel® architecture optimizations using Intel Distribution of OpenVINO™ toolkit and quantization with Vector Neural Network Instructions (VNNI) improve performance 8.24X[1]

INTEL®
AI BUILDERS
MEMBER

汇医慧影
huiyihuiying.com

According to the National Cancer Institute, about one in eight women will be diagnosed with breast cancer at some point in their lives.[2] It is the most commonly diagnosed cancer[3] and a leading cause of cancer death in women.[4]

Ultrasound, mammography, MRI, and other imaging techniques are important means of diagnosis for breast cancer. However, in the field of medical imaging, there is a gap between doctor supply and demand, high barriers to entry, and uneven distribution of medical resources around the world. Yet healthcare imaging still contributes to the rapid growth of data in hospital and clinical settings, with not enough doctors to quickly provide analysis and diagnosis.

"Doctors often deal with over 1,000 medical image films daily, which would make even the most experienced doctors fatigued in an overwhelming ocean of data," said Yonggao Zhang, Deputy Director, Radiology, The First Affiliated Hospital of Zhengzhou University in Zhengzhou City, China.

The growing repositories of patient imaging data, a shortage of radiologists, and advances in computing technologies are driving companies like Huiying Medical Technology to provide solutions that utilize artificial intelligence (AI) to help clinicians quickly and efficiently analyze imagery and diagnose critical cases with high accuracy.

A leading developer of healthcare technology solutions in China, Huiying Medical develops AI-driven medical imaging analytics systems to aid early screening and diagnosis for a range of medical conditions, including lung nodules, tuberculosis (TB), breast cancer, bone fractures, and others. Huiying Medical Technology is dedicated to applying computer vision and deep learning to the medical field. The company's suite of medical diagnosis and medical data analysis solutions are based on cloud computing, big data, and AI.

Whole-Cycle Breast Cancer Detection and Clinical Decision-Making Support Solution

Today's AI-based imaging diagnosis support systems are often limited to single screening applications for triage or pre-screening before a doctor's analysis and diagnosis. However, during their disease cycle, patients might undergo many treatment protocols and be seen by more than one physician over a long time period. Huiying's Dr. Turing AI-Assisted Diagnosis Platform—Breast Cancer Detection Solution runs through the patient's entire disease cycle with capabilities to assist clinicians in diagnosis and consistently communicate a patient's risk of developing breast cancer through the full monitoring period. These include:

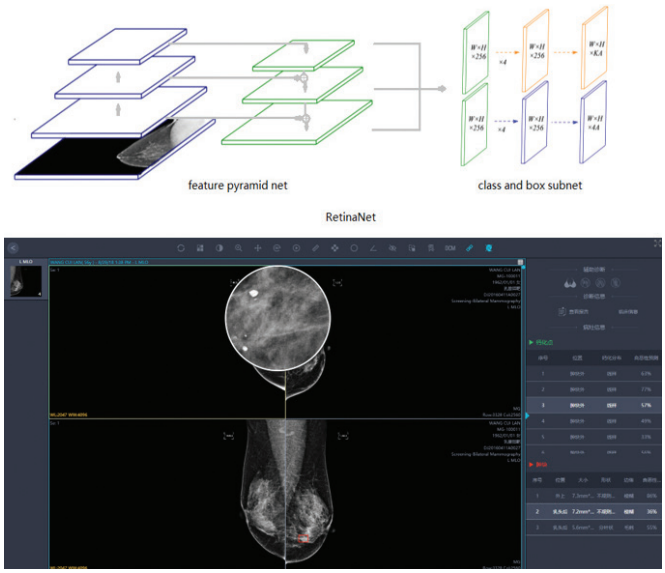
- Accelerated image analysis assistance
- Structured image reporting based on American College of Radiology (ACR) standards
- Automatic updating of patient information in the Breast Imaging-Reporting and Data System (BI-RADS)

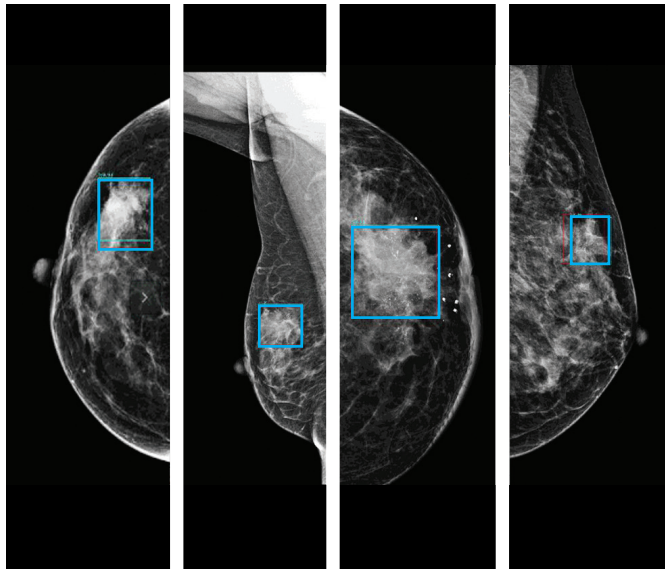
Huiying's solution is a next-generation, AI-based intelligent image diagnosis and multi-modality system to help medical staff improve efficiency in imaging, diagnosis, clinical decision-making, and disease management. The technology is designed for both clinical accuracy and efficiency, which is expected to help reduce cost of diagnosis and treatment.

Leveraging AI and Intel Technologies

Working with Intel, Huiying Medical's data scientists, engineers, and software developers optimized their model training and inferencing of breast cancer detection and diagnosis technologies. Huiying's breast cancer screening technology utilizes deep neural networks, such as Inception V4 and Inception ResNet V2, to support multi-modality data operations and greatly improve data processing and inference efficiency. Their system was built using the TensorFlow framework, the Intel® Distribution of OpenVINO™ toolkit, and PyTorch deep learning library and based on RetinaNet with Resnet50 as the Convolutional Neural Network (CNN) backbone. The Intel Distribution of OpenVINO toolkit is a free software kit that helps developers and data scientists speed up AI inferencing workloads and streamline deep learning deployments from the network edge to the cloud across Intel architectures—CPUs, integrated GPUs, ASICs, and FPGAs.

Optimized to run on 2nd Generation Intel Xeon® Scalable processors, Huiying adopted the Intel Distribution of OpenVINO toolkit to accelerate performance of computer vision workloads, as well as Intel Optimizations for TensorFlow, the Intel Math Kernel Library for Deep Neural Networks (Intel MKL-DNN), and the Vector Neural Network Instruction set (VNNI) to accelerate inferencing.





8.24X Faster Inferencing with Optimizations and INT8 Quantization

The original model for Huiying's breast cancer detection and diagnosis solution was based on a Keras FP32 model. The Inference Engine inside the Intel Distribution of OpenVINO toolkit enables deployment of network models. In order to perform the inference, the Inference Engine operates with an intermediate representation of the original model using the Model Optimizer tool, which is optimized for execution on end-point target devices. The model is then quantized to a lower precision—from FP32 to INT8—which takes up less memory footprint and accelerates inference performance when using Intel Deep Learning Boost. Intel Deep Learning Boost is built into 2nd Gen Intel Xeon Scalable processors.

Running on 48-core Intel architecture-based platforms with two Intel Xeon 8268 processors per server, the software could execute multiple operations in parallel. This resulted in much faster compute times, especially on inference-based tasks, like processing several streams of medical scans using Intel Deep Learning Boost.

As shown in Table 1 and Figure 1, the higher precision FP32 model, leveraging the Intel Distribution of OpenVINO toolkit, improved performance 3.02X[5,6] over the original Keras model. INT8 quantization with Intel DL Boost further improved performance to 8.24X over the original model. Accuracy loss was less than -0.17 percent. The ability to lower the precision of a model from FP32 to INT8 is a great practice to accelerate the performance of certain models on hardware that supports INT8. An INT8 model takes up less memory footprint and speeds up inference time with a small reduction in accuracy.

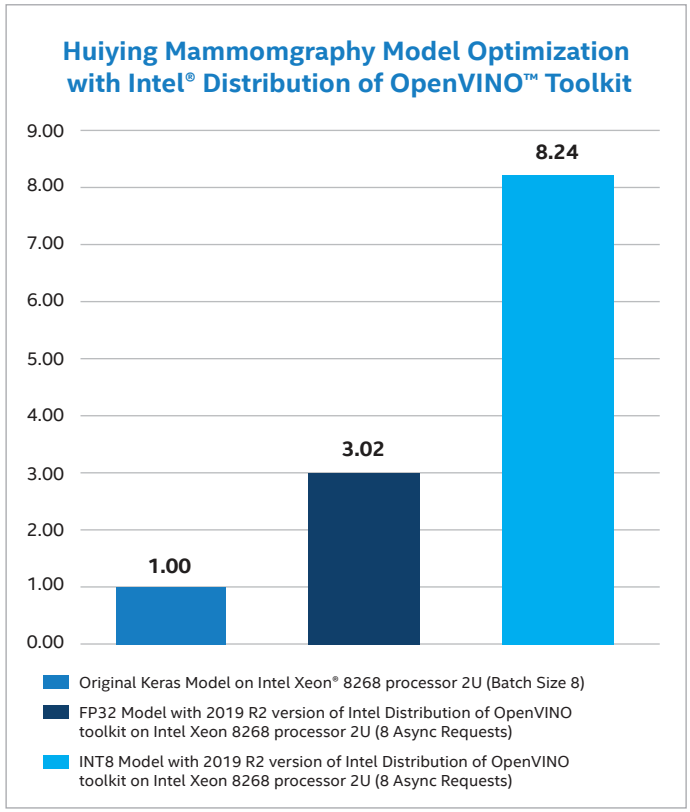


Figure 1. Intel optimization performance results on Huiying Breast Cancer Solution Models

Table 1. Intel optimization performance results on Huiying Breast Cancer Solution Models

	CONFIGURATION 1	CONFIGURATION 2	CONFIGURATION 3
Throughput (frames/sec)	3.4	10.4	28.3
Latency (ms)	291	96.2	35.3
Speedup (Throughput/Latency) Compared to Baseline		3.02X	8.24X
Accuracy Loss—INT8 vs. FP32			-0.17%

Testing Overview (see notes for details):

- Model: RetinaNet-based breast cancer detection model
- Dataset: Customer dataset with 366 mammography images. Model Input Size: 1280X640
- Configuration 1—Baseline: Original Keras model, BS=8 on two sockets
- Configuration 2—FP32 Performance: BS=1, OpenVINO, 8 async instanced on two sockets
- Configuration 3—INT8 Performance: BS=1, OpenVINO, VNNI, 8 async instances on two sockets

Conclusion

The imbalance between demand and supply of healthcare imaging and clinical resources the world is seeing makes early detection and intervention of cancer extremely important. Early detection saves lives and helps reduce costs in healthcare. Optimized on Intel technologies, the Dr. Turing AI-Assisted Diagnosis Platform—Breast Cancer Detection Solution, accelerates image analytics performance up to 8.24X. With AI-supported diagnosis, the solution helps reduce false positives, patient recalls, and unnecessary biopsies for both mass and calcifications.

Designed to support monitoring patients throughout the entire disease cycle, the system also provides automation of standardized communication and reporting for the multiple physicians that a patient might see. These improvements ease the burden on radiologists while increasing the level of care they can provide.

For more information about the Huiying platform, visit https://huiyihuiying.com.

INTEL® AI BUILDERS MEMBER

Huiying Medical Technology is a member of the **Intel AI Builders Program**, an ecosystem of industry-leading independent software vendors (ISVs), system integrators (SIs), original equipment manufacturers (OEMs), and enterprise end users, which have a shared mission to accelerate the adoption of artificial intelligence across Intel platforms.

1. Tested by Intel as of 9/18/2019. 2 socket Intel® Xeon® Platinum 8268 Processor, 28 cores HT On Turbo ON Total Memory 384 GB (12 slots/32 GB/2933 MHz), BIOS: SE5C620.86B.0X.02.0001.051420190324 (ucode:0x4000024), CentOS Linux* 7 (Core)), Intel Software: OpenVINO 2019.2.275, Topology: RetinaNet: https://github.com/fizyr/keras-retinanet, Compiler: gcc 4.8.5,MKL DNN version: v0.17, BS=1, 8 Asynchronous Requests, customer data, 1 instance/2 socket , Datatype: Int8
2. https://www.cancer.gov/types/breast/risk-fact-sheet
3. https://www.wcrf.org/dietandcancer/cancer-trends/worldwide-cancer-data
4. https://www.cdc.gov/cancer/breast/statistics/index.htm
5. Keras baseline model: Tested by Intel as of 9/18/2019. 2 socket Intel® Xeon® Platinum 8268 Processor, 28 cores HT On Turbo ON Total Memory 384 GB (12 slots/ 32GB/ 2933 MHz), BIOS: SE5C620.86B.0X.02.0001.051420190324 (ucode:0x4000024), CentOS Linux* 7 (Core)), Deep Learning Framework: Keras 2.2.4 and Intel-TensorFlow: 1.13.1, Topology: RetinaNet: https://github.com/fizyr/keras-retinanet, Compiler: gcc 4.8.5,MKL DNN version: v0.17, BS=8, customer data, 1 instance/2 socket, Datatype: FP32
6. Tested by Intel as of 9/18/2019. 2 socket Intel® Xeon® Platinum 8268 Processor, 28 cores HT On Turbo ON Total Memory 384 GB (12 slots/ 32GB/ 2933 MHz), BIOS: SE5C620.86B.0X.02.0001.051420190324 (ucode:0x4000024), CentOS Linux* 7 (Core), Intel Software: OpenVINO™ 2019.2.275, Topology: RetinaNet: https://github.com/fizyr/keras-retinanet, Compiler: gcc 4.8.5,MKL DNN version: v0.17, BS= 1, customer data, 8 Asynchronous Requests, 1 instance/2 socket, Datatype: FP32

Performance results are based on testing as of dates shown in configuration and may not reflect all publicly available security updates. No product or component can be absolutely secure. See configuration disclosure for details.

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. 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: http://www.intel.com/performance

 1119/AU/HBD/PDF ♻ Please Recycle 341779-001US