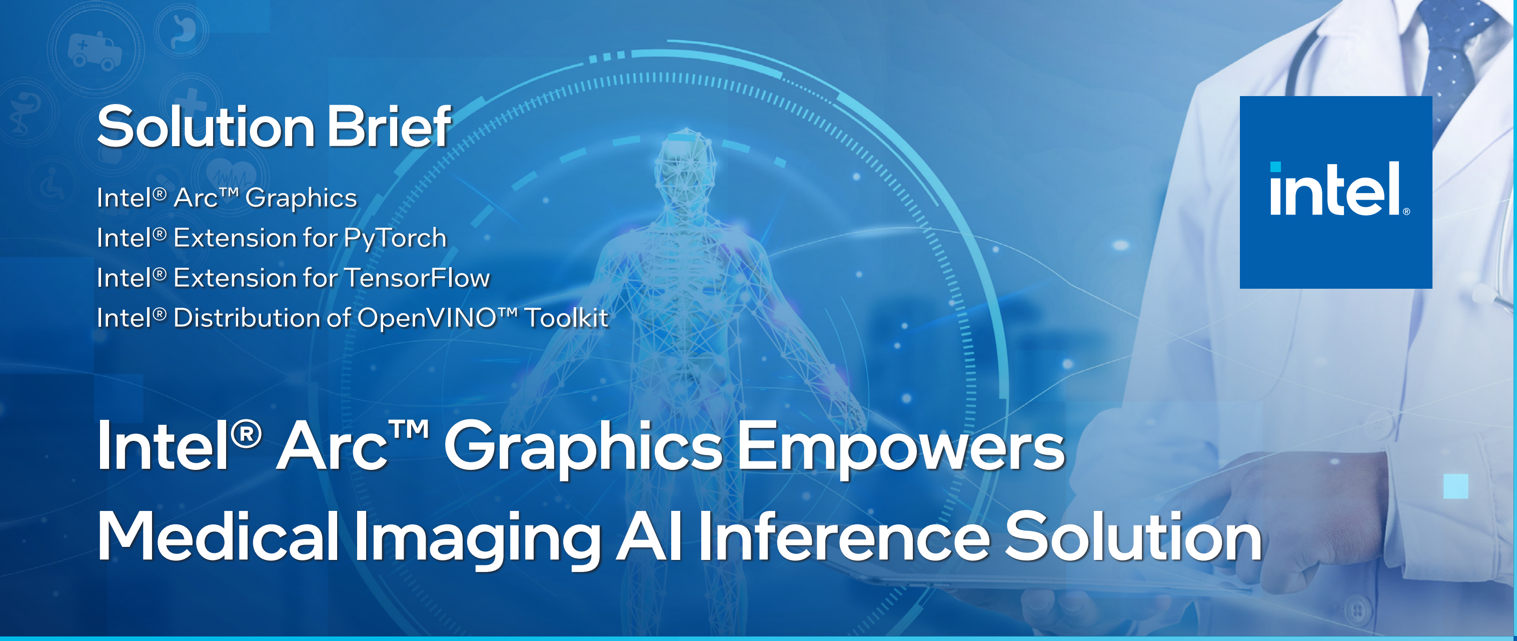


Solution Brief

Intel® Arc™ Graphics
Intel® Extension for PyTorch
Intel® Extension for TensorFlow
Intel® Distribution of OpenVINO™ Toolkit

Intel® Arc™ Graphics Empowers Medical Imaging AI Inference Solution

Overview

As algorithms become increasingly mature, computing power increases, and data continue to grow exponentially, artificial intelligence (AI) has been broadening its application scope in medical imaging. The medical imaging AI applications can effectively improve the diagnostic efficiency and accuracy, bring medical imaging analysis technology to frontline medical staff, and shorten the waiting time of patients to provide them better services. Meanwhile, however, medical imaging AI applications also face challenges such as difficult access to computing resources, high cost, limited choices of computing resources, high computing demand, and needs of medical hardware customization.

To help users address the above challenges, Intel has launched an edge AI inference solution based on Intel Arc™ Graphics. The solution gives full play to the potential of Intel Arc™ Graphics in AI inference, and accelerates algorithm migration and AI inference with the help of Intel® Extension for PyTorch (IPEX), Intel® Extension for TensorFlow (ITEX), OpenVINO™ Toolkit, and Intel® oneAPI Base Toolkit. With low TCO, rich product line, high compatibility, and a variety of custom hardware for medical use, the solution helps partners quickly build excellent medical imaging AI inference solutions, which satisfy key performance indicators such as AI-assisted medical imaging diagnosis and treatment.

Background: Edge AI enables smart hospital

AI is widely used in healthcare and playing a critical role in many aspects from medical imaging, AI-assisted diagnosis, disease prediction, and health management, to drug research and development. With the progress of deep learning technology, AI can help applications generalize relevant features from massive data automatically, instead of manually discovering and designing features according to domain-specific knowledge like traditional models. This enables users to train high-quality AI medical models more quickly and respond to the AI-assisted diagnosis needs of different scenarios more flexibly.

With the continuous penetration of AI in healthcare, edge computing has developed rapidly. With the help of edge terminals deployed near data sources that integrate network, computing, storage and applications, the edge AI can transfer the inference part of AI workflow from the cloud or data center to the nearest edge for computing, thereby reducing delay, saving network bandwidth, and protecting privacy. As a result, edge AI applications have been widely deployed in various medical fields in recent years. Edge AI applications provide doctors with reference, increase the number of cases doctors are able to handle, improve the accuracy of image analysis, and shorten the time of diagnosis report. This is of great significance for primary-level medical institutions to enhance the diagnosis and treatment capacity and promote disease prevention and treatment.

To promote the deployment of edge AI systems in healthcare, the stringent requirements of AI model inference for computing power should be satisfied first. Meanwhile, the edge AI systems for the healthcare industry need to address many challenges in stability and economy:

- **Difficult access to computing resources and high deployment cost:** To achieve higher flexibility and cost-effectiveness and avoid binding computing resources to a single vendor, users generally hope to have more options of inference computing.
- **High computing resources requirements:** Medical AI applications rely on the inference of deep learning models. With the improvement of model complexity and the growth of data volume, the system inference capability faces huge challenges. In addition, as customers hope to have higher user concurrency and meet the detection time of target applications, such challenges grow bigger.
- **Need of custom hardware more suitable for medical scenarios:** As medical imaging AI is increasingly applied to clinical scenarios, the AI inference part needs to be transferred to the edge, raising higher requirements for edge devices. Medical edge AI terminals need to be used in demanding medical scenarios, which require low noise, high stability, waterproof and antibacterial, and medical certification.

Medical imaging AI inference solution based on Intel® Arc™ Graphics

Based on the long-term technical expertise and innovation in GPU, as well as its active GPU ecosystem, Intel has launched Intel® Arc™ Graphics. This series of graphics have powerful computing, rich software functions, and an ecosystem that supports algorithms compatibility and migration, all of which accelerate the implementation of AI applications in various industries at the edge. In the healthcare industry, Intel provides a medical imaging AI inference solution for reference based on Intel® Arc™ Graphics, which helps independent software developers (ISV), original equipment manufacturers (OEM) and other partners develop their high-performance, cost-effective, and highly flexible medical imaging AI inference solutions at the edge.

As shown in Figure 1, the solution's underlying hardware is Intel® CPU and Intel® GPU. With Intel® Extension for PyTorch, Intel® Extension for TensorFlow, OpenVINO™ tool suite and Intel® oneAPI toolkit, it migrates AI algorithms, optimizes algorithm performance, and supports the efficient operation of AI algorithms in the healthcare industry.

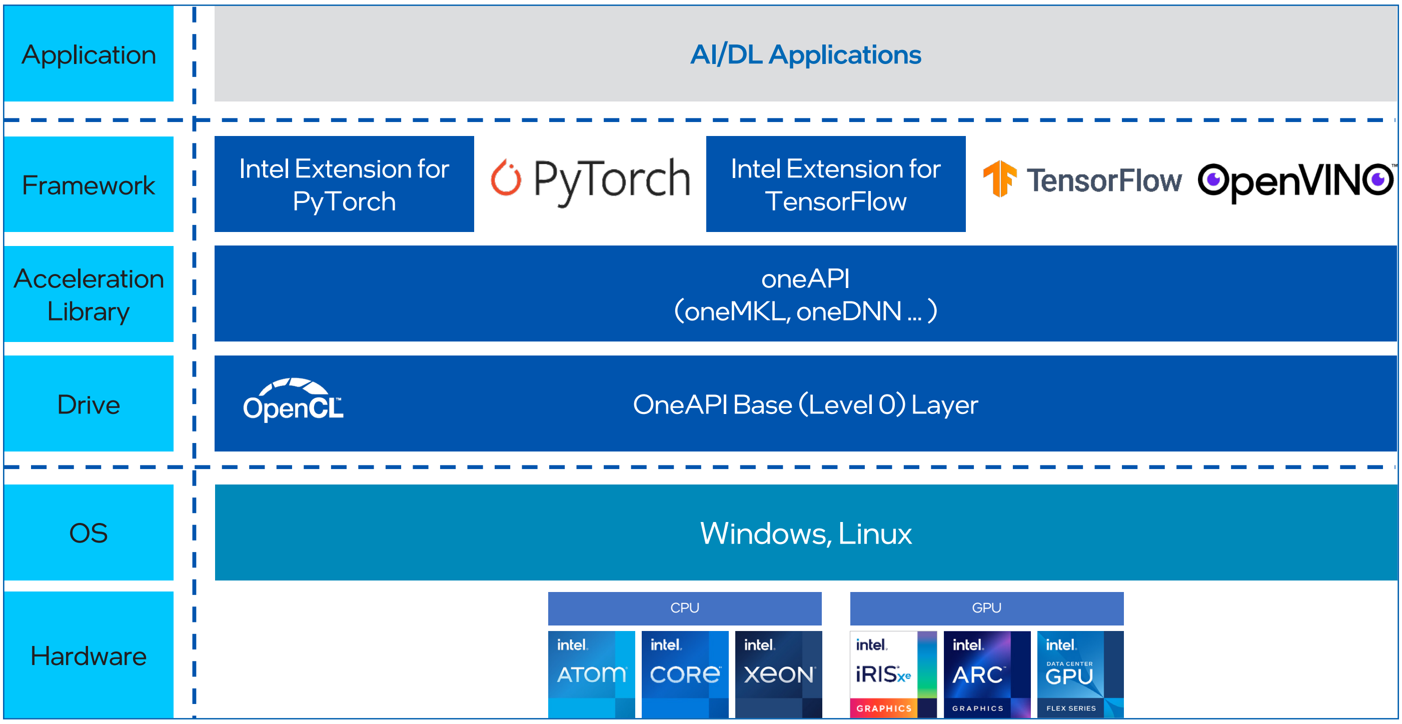

Application
AI/DL Applications
Framework
Intel Extension for PyTorch
PyTorch
Intel Extension for TensorFlow
TensorFlow
OpenVINO
Acceleration Library
oneAPI
(oneMKL, oneDNN ...)
Drive
OpenCL
OneAPI Base (Level 0) Layer
OS
Windows, Linux
Hardware
CPU
intel ATOM
intel CORE
intel XEON
GPU
intel iRIS Xe GRAPHICS
intel ARC GRAPHICS
intel DATA CENTER GPU FLEX SERIES

Figure 1. Reference architecture of medical imaging AI inference solution based on Intel® Arc™ Graphics

High-performance hardware base

Intel® Arc™ Graphics and Intel® Core™ processors are recommended for the medical imaging AI inference solution. Among them, Intel® Arc™ Graphics have up to 32 X^e cores, 8 rendering slices, a maximum core frequency of 2400MHz, and a maximum graphics memory of 16 GB 256bit GDDR6. The X^e kernel of Intel® Arc™ Graphics integrates Extended Vector Engine (XVE) and Extended Matrix Engine (XMX), which accelerate AI workflow and provide powerful and real-time computing power support for AI inference at the edge.

Figure 2. Intel® Arc™ Graphics

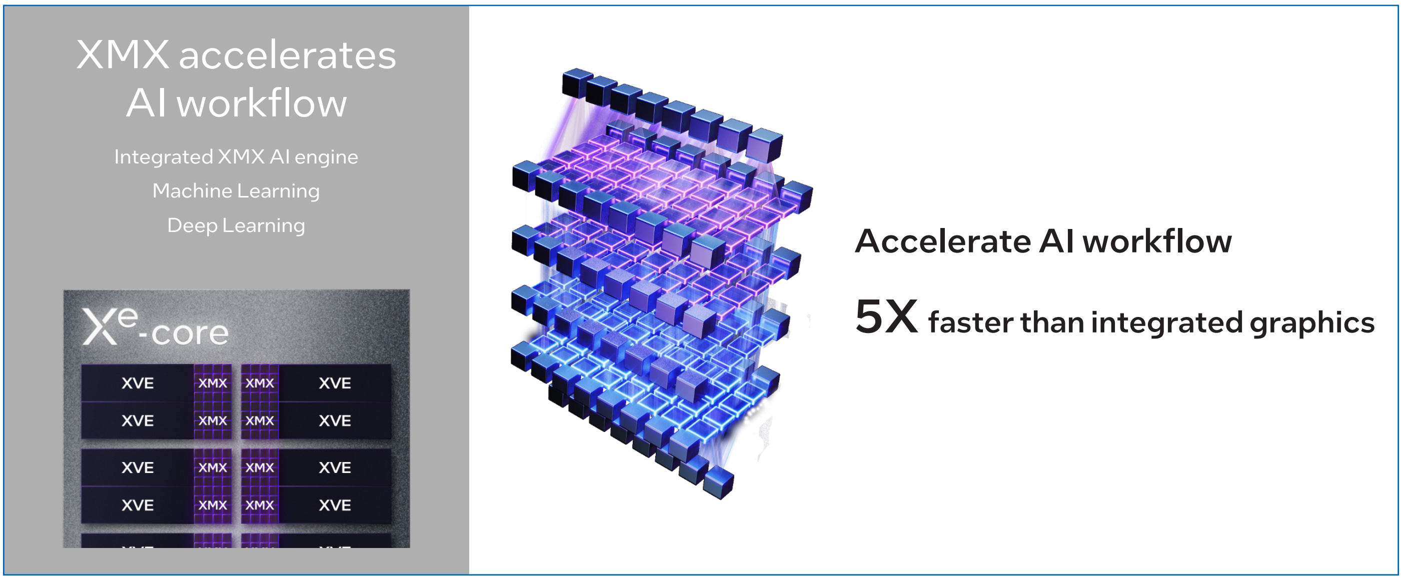


Figure 3. Intel® Arc™ Graphics deliver excellent performance[1]

The 13th Gen Intel® Core™ processors redefine the multicore architecture for edge and end devices with a brand-new performance hybrid architecture. The P-cores (Performance) significantly improve single-thread performance and response speed, and E-cores (Efficiency) support multitasking with scalable multithread performance and efficient background task offloading. The processor includes Intel® Arc™ X^{e} Graphics powered by Intel® X^{e} architecture, and up to 96 execution units (EUs) that can perform computing tasks in parallel with the Intel® Arc™ Graphics.

AI algorithm ecosystem with easy compatibility and migration

With the help of Intel® Extension for PyTorch (IPEX), Intel® Extension for TensorFlow (ITEX), OpenVINO™ Toolkit, and Intel® oneAPI Base Toolkit, the medical imaging AI inference solution based on Intel® Arc™ Graphics can facilitate the migration of AI algorithms and optimize performance for Intel® hardware.

● Using Intel® Extension for PyTorch to migrate PyTorch-based models

For models written based on PyTorch, in order to achieve algorithm migration and additional performance improvements on Intel® hardware, partners can get support with Intel® Extension for PyTorch optimized for GPUs. Intel® Extension for PyTorch is an open-source extension initiated by Intel based on the extension mechanism of PyTorch. It gives full play to the hardware features by offering additional software optimizations, in order to help users, on the basis of native PyTorch, significantly improve deep learning inference and training performance on Intel® hardware (such as CPU and GPU). With the extension, PyTorch users will be able to take full advantage of the latest features of Intel hardware and enjoy the superior performance and deployment convenience brought by software optimizations as early as possible.

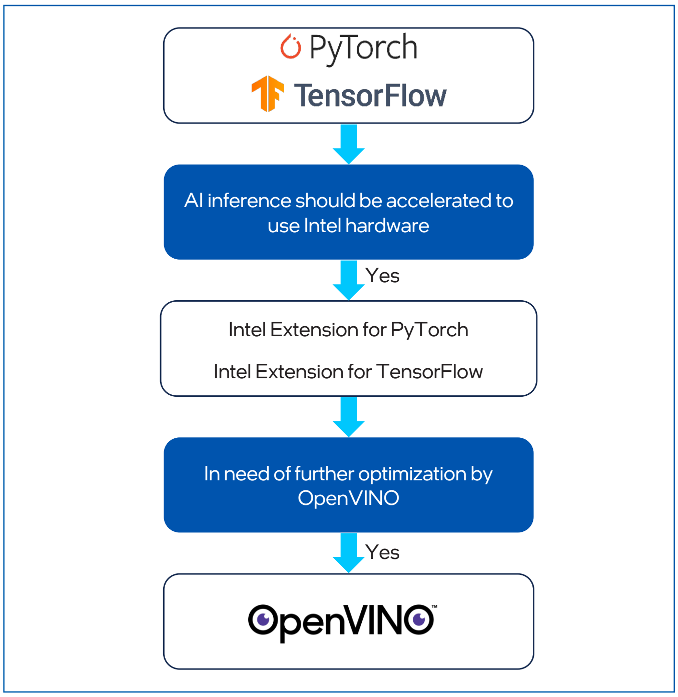

PyTorch
TensorFlow
AI inference should be accelerated to use Intel hardware
Yes
Intel Extension for PyTorch
Intel Extension for TensorFlow
In need of further optimization by OpenVINO
Yes
OpenVINO

Figure 4. Intel software supports easy model migration and acceleration

[1] Intel technologies may require enabled hardware, software or service activation. Performance varies by use, configuration and other factors. For more information, please visit www.Intel.com/PerformanceIndex

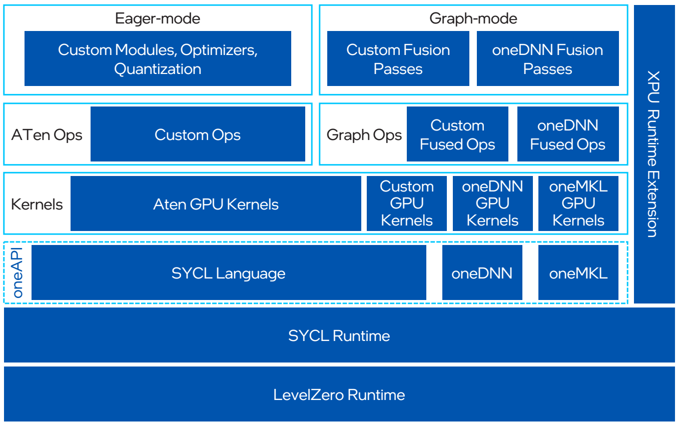


Figure 5. Framework of Intel® Extension for PyTorch

Intel® Extension for PyTorch (IPEX) supports Intel® GPUs and enables the latest AI optimizations of Intel software and hardware on Intel® GPUs, including uploading multiple optimizations to the inventory versions of framework for out-of-the-box performance improvements. The additional performance of IPEX comes from the optimization for eager-mode and graph-mode. In the eager-mode, PyTorch frontend is extended with custom Python modules (such as fusion modules), better optimizers, and INT8 quantization APIs. Further performance improvement is achieved by converting eager-mode models into graph-mode using extended graph fusion passes. On GPU, optimized operators and kernels are implemented and registered through PyTorch dispatching mechanism.

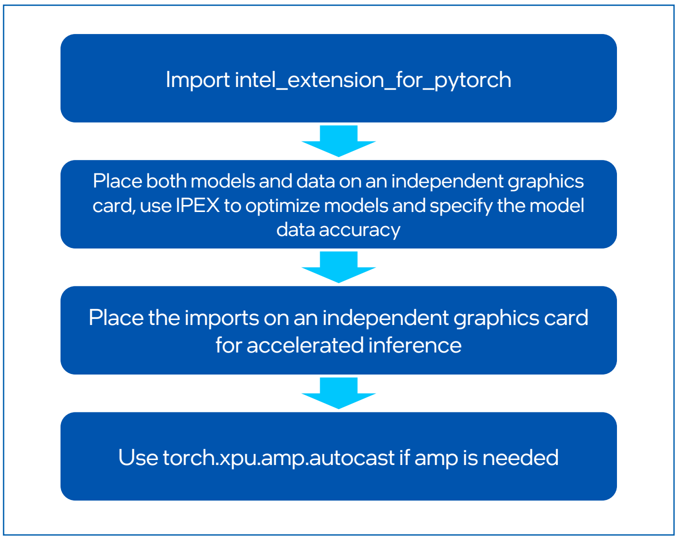
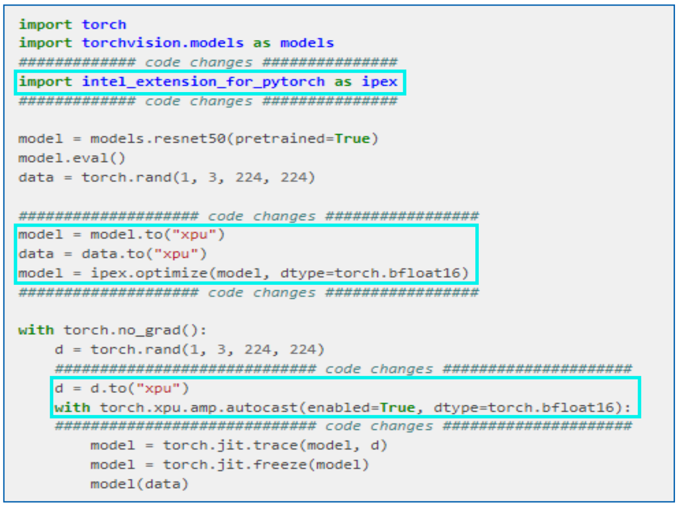

Import intel_extension_for_pytorch
Place both models and data on an independent graphics card, use IPEX to optimize models and specify the model data accuracy
Place the imports on an independent graphics card for accelerated inference
Use torch.xpu.amp.autocast if amp is needed

Figure 6. The flowchart of applications migration using Intel® Extension for PyTorch

The example of code changes by importing Intel® Extension for PyTorch is shown in Figure 6. With Intel® Extension for PyTorch, users can quickly migrate codes to run on diverse independent GPUs without tedious code development. The extension can also be loaded as a C++ library for C++ programs.



```
import torch
import torchvision.models as models
############# code changes ###############
import intel_extension_for_pytorch as ipex
############# code changes ###############

model = models.resnet50(pretrained=True)
model.eval()
data = torch.rand(1, 3, 224, 224)

#################### code changes #################
model = model.to("xpu")
data = data.to("xpu")
model = ipex.optimize(model, dtype=torch.bfloat16)
#################### code changes #################

with torch.no_grad():
    d = torch.rand(1, 3, 224, 224)
    ############################# code changes #####################
    d = d.to("xpu")
    with torch.xpu.amp.autocast(enabled=True, dtype=torch.bfloat16):
    ############################# code changes #####################
        model = torch.jit.trace(model, d)
        model = torch.jit.freeze(model)
        model(data)
```

Figure 7. Example of code changes[2]

● Using Intel® Extension for TensorFlow to migrate TensorFlow-based models

For models written based on TensorFlow, partners can migrate them using Intel® Extension for TensorFlow[3]. Intel® Extension for TensorFlow is a high performance deep learning extension plugin based on TensorFlow PluggableDevice interface. Through seamless integration with the TensorFlow framework, it allows TensorFlow developers to have an easy access to Intel GPUs and CPUs. With this Intel extension, developers can train TensorFlow models for inference on Intel AI hardware with zero code changes.

The Intel® Extension for TensorFlow is built on the oneAPI software component. Most performance-critical diagrams and operators are highly optimized with Intel® oneAPI Deep Neural Network Library (oneDNN), which is an open source, cross-platform library that provides implementations of deep learning building blocks. Other operators are implementations of SYCL, a core API language for programming accelerators and multiprocessors.

[2] For more examples of code changes, please visit https://github.com/intel/intel-extension-for-pytorch/blob/v2.0.110%2Bxpu/examples/gpu/inference/python/resnet50_torchscript_mode_inference_bf16.py

[3] For more information, please visit https://www.intel.com/content/www/us/en/developer/articles/technical/introduction-to-intel-extension-for-tensorflow.html

● Using OpenVINO™ Toolkit to migrate models based on ONNX and PaddlePaddle

For models written based on frameworks like ONNX and PaddlePaddle, OpenVINO™ Toolkit can be used for their migration[4]. OpenVINO™ toolkit is a comprehensive suite of toolkit for quick development of applications and solutions to solve a variety of tasks, including human visual simulation, automatic speech recognition, natural language processing, and recommendation systems. The OpenVINO™ toolkit provides direct support for models based on frameworks like ONNX or PaddlePaddle, and the Model Optimizer can directly complete the offline migration of the models mentioned above.

OpenVINO™ Toolkit also brings a high improvement on AI inference performance to models. Based on the latest generation of artificial neural networks, including convolutional neural networks (CNNs), recurrent networks, and attention-based networks, this toolkit scales computer vision and non-visual workloads across Intel® hardware, thereby improving performance significantly. It accelerates applications with high-performance AI and deep-learning inference deployed from edge to cloud.

For models written based on other frameworks, Intel also offers Intel® oneAPI Base Toolkit[5] that helps quickly complete the migration. Intel® oneAPI Base Toolkit is an Intel® software development tool based on new generation standards, aiming to build and deploy data-centric high-performance applications across a variety of architectures. It accelerates the computing process by leveraging excellent hardware features, and is fully compatible with existing programming models and codebases, ensuring that applications written by developers can run seamlessly on oneAPI.

The Intel® oneAPI Base Toolkit can automatically migrate the codes written based on other frameworks to Intel® Arc™ Graphics. Through this migration, users can reduce the complexity of cross-platform development and migration of AI inference tasks, improve the performance of AI models running on heterogeneous platforms, and make full use of existing medical AI models, thus accelerating the development of medical AI applications.

In addition, this reference solution also provides rich software functions, which help medical institutions efficiently process medical imaging data to meet the needs of advanced smart medical scenarios. For example, Intel provides an open source OpenGL driver, which enables medical institutions to address the needs of 3D visualization processing like 3D reconstruction in medical imaging-assisted diagnosis, so that doctors can have a clear view of the spatial relationship between body structures from multiple angles.

Custom hardware for medical scenarios at the edge

With the proliferation of medical imaging AI in clinical applications and intelligence in medical specialty, AI inference is often integrated with edge computing, which is different from the conventional loose coupling of server and software. The hardware design is more strictly required to meet the use scenarios of medical staff. For example, AI terminals need to be used in demanding medical scenarios, which require low noise, high stability, dustproof, waterproof, and antibacterial, as well as safety regulations and EMC certification.

Typical medical terminals based on Intel® architecture adopt outstanding cooling solutions, which feature large heat emission holes, fans, and water cooling to ensure CPU and GPU performance, and maintain noise at a low level to ensure a quiet medical office environment. These terminals offer personalized customization to help customers differentiate products, and support AI-assisted diagnosis platforms and various add-on card functions, thus effectively supporting the AI computing requirements of AI-assisted diagnosis software deployed at the local end. These terminals also have rich I/O ports, which can comprehensively further the smart AI healthcare goal with easy connection to various devices and convenient deployment.

Figure 8. Intel® architecture-based medical terminals meet various requirements

[4] For more information, please visit https://www.intel.cn/content/www/cn/zh/developer/tools/openvino-toolkit/overview.html

[5] For more information, please visit https://www.intel.com/content/www/us/en/docs/dpcpp-compatibility-tool/get-started-guide/2023-0/overview.html

Benefits

The medical imaging AI inference solution based on Intel® Arc™ Graphics leverages the integration of Intel hardware and software, supporting efficient and stable operation of AI applications at the edge.

For partners, this solution can bring the following important values:

- **Reduced TCO:** The solution makes full use of the advantages of Intel® CPU and GPU, as well as Intel® software and technology, to improve cost performance and promote the popularity of medical imaging AI applications.
- **Extensive product line:** Intel provides a complete product line covering CPU, GPU and software tools for more convenient solution construction.
- **High compatibility:** The software suite offered by Intel can flexibly extend across multiple Intel® hardware, and run in parallel with integrated graphics through independent graphics to further improve performance. With this solution, ISV and OEM will have more options for inference computing, avoiding the risks of relying on a single device offered by a single supplier.
- **Various custom hardware:** Intel has a large number of hardware partners. They provide a variety of custom hardware to meet requirements such as real-time, reliability, low noise, waterproof, antibacterial, and medical certification.

Application Practice

So far, a number of partners have introduced corresponding products and solutions by referring to the medical imaging AI solution based on Intel® Arc™ Graphics. For example, HY Medical has introduced an AI-assisted bone mineral density (BMD) detection system based on CT scanning images of thorax and abdomen. It automatically measures and analyzes the BMD, and simultaneously presents vertebral analysis data, intervertebral analysis data, and abdominal tissue composition analysis data.

To verify the performance of this solution, HY Medical ran some tests. The BMD screening project included 4 AI models: a coarse spine segmentation model based on 3D-Resnet-Unet, a spine segmentation model based on 3D-Unet, a BMD regression model based on 3D-DenseNet, and a tissue segmentation model based on 2D-Unet. The test data come from a total of 120 cases of CT scanning on body parts including chest, abdomen, and chest-abdomen. The diagram of typical test samples is shown in Figure 9:

The test results show that the average time of AI inference based on images only is 9.55 seconds[6]. This indicates that with such superior performance, Intel® Arc™ Graphics is capable of helping the detection system efficiently measure and analyze the BMD data.

"With the help of AI screening, we can proactively remind patients of possible bone density problems and prevent them early." Chai Xiangfei, CEO of HY Medical, pointed out, "As we built our AI solution to BMD screening, Intel® Arc™ Graphics provided powerful computing capability to help us meet performance goals easily. It also excels in flexibility and stability, which will help drive large-scale application of BMD screening."

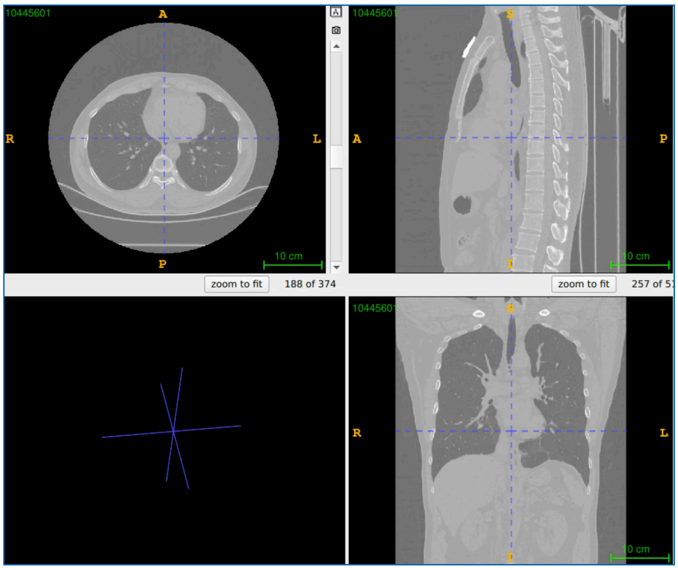


Figure 9. Diagram of typical BMD test samples

[6] Data quoted from internal test results of HY Medical as of August 2023. Test configurations: Intel® Core™ i9-12900K processor (12 cores, 24 threads), 64 GB total memory, Intel® Arc™ Graphics (16G). Intel technologies may require enabled hardware, software or service activation. Intel does not control or audit third-party data. You should consult other sources to evaluate accuracy.

Prospects

The *Outline of the Healthy China 2030 Plan* proposes that to fulfill the fundamental aim of "health for all", the whole healthcare life cycle must be covered, strengthening intervention in major health problems and influence factors at different life stages, and delivering whole-process health services and security. The medical imaging AI inference solution based on Intel® Arc™ Graphics helps hospitals improve medical images reading capability, reduce their requirements for resources like doctors and devices, and improve the efficiency of disease diagnosis, giving patients advice on timely appropriate countermeasures.

Intel will strengthen innovation in key technologies like GPU and enhance ecosystem cooperation. On the one hand, Intel will use more optimization strategies, such as parallel running of independent graphics and integrated graphics, and continuous optimization of solution performance. On the other hand, Intel will find more ways of applying Intel® Arc™ Graphics in multiple AI medical fields, so as to unleash the potential of Intel® Arc™ Graphics in AI acceleration in a wider range of scenarios, and help to drive the transformation of smart hospitals.

Disclaimers: This article is intended solely to promote Intel and partner technology. Intel does not in any way promote or introduce medical institutions, medical services, nor does it recommend or certify any drugs, medical devices, health food, etc.

Performance varies by use, configuration and other factors. Learn more at www.Intel.com/PerformanceIndex

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

Your costs and results may vary.

Intel technologies may require enabled hardware, software or service activation.

Intel disclaims all express and implied warranties, including without limitation, the implied warranties of merchantability, fitness for a particular purpose, and non-infringement, as well as any warranty arising from course of performance, course of dealing, or usage in trade.

Intel does not control or audit third-party data. You should consult other sources to evaluate accuracy.