Intel China

2020 AI Implementation Guide for the Healthcare Industry

contents

Trends

Implementation

Technologies

Hardware product

Software and framework

Note: The sections marked with * are 2020 updates

Trends

Development and Application of AI in Healthcare

Development of AI in Healthcare

Market Trend of AI in Healthcare

Thanks to the further innovations of algorithms, the improvement of computing power and the continuous growth of data, Artificial Intelligence (AI) is undergoing rapid development, especially in the field of deep learning, and its application increasingly focuses on certain industries. In September 2018, the ***2018 Blue Paper on World AI Industry*** released by China Academy of Information and Communications Technology pointed out that, among all kinds of vertical industries in China, the penetration rate of AI in Healthcare, Finance, Commerce, Education, Security, etc. is relatively high and AI companies specializing in healthcare represent the highest percentage (22%) [1].

The healthcare industry, as one of the key application fields of AI, is also investing more and more money into AI technologies. According to the ***Analysis Report on the Market Outlook and Investment Strategy Planning of China's AI Industry from 2018 to 2023*** released by Outlook Industry Research Institute, the market size of China's medical AI industry reached 9.661 and 13 billion yuan in 2016 and 2017, with an increase of 37.9% and 40.7% respectively. In 2018, the market size was expected to reach 20 billion yuan[2]. This rapid growth benefited from the urgent needs of China's medical market, the development of medical AI technologies as well as relevant support policies in recent years.

Globally, the market segmentation of medical AI application is slightly different from that in China. According to statistics from Global Market Insight, drug R&D accounts for the largest share of the global medical AI market, reaching 35%. This is closely followed by medical imaging AI (accounting for 25%), which will grow by more than 40% and is expected to reach 2.5 billion US dollars in 2024. [3]

In addition, genomics analysis is another important field of AI application. It is estimated that by 2022, the size of this market segment will reach approx. 30 billion yuan in China alone[4]. The further integration with AI will accelerate the development of gene sequencing and shorten the sequencing time and greatly reduce the sequencing cost, which will expand the vision for AI application in the healthcare industry.

It is worth mentioning that AI is playing an important role in the prevention and control of the COVID-19 pandemic, and has already shown its unique capabilities in intelligent service robots, intelligent big data analysis, temperature check, auxiliary diagnosis, genome sequencing, drug R&D, etc., further demonstrating the broad prospects of AI application in the healthcare industry.

In China, government incentives are one of the key factors that accelerate the application of medical AI. Since 2015, relevant government authorities have successively introduced nearly 20 policies covering talent training, technological innovation, standard and supervision, industry integration, product and application, etc., to promote the development of AI. In March 2017, the word "AI", for the first time, appeared in the Work Report of the Chinese Government. In July 2017, the State Council issued the ***Development Plan for the New Generation of AI***, clearly setting up a 3-phase strategy for the development of new generation of AI. In October 2017, AI was referenced in the Report of the 19th CPC National Congress in which the "deep integration of Internet, Big Data, and AI with real economy" were identified as the development direction of China's digital economy. In December 2017, the Ministry of Industry and Information Technology released the ***Three-Year Action Plan for Promoting the Development of New Generation AI Industry (2018-2020)***, which detailed the key development directions and targets of AI in the next three years. In January 2018, the ***White Paper on Standardization of AI (2018 Edition)*** was released under the guidance of the National Standardization Administration. In April 2018, the State Council issued the ***Opinions on Promoting the Development***

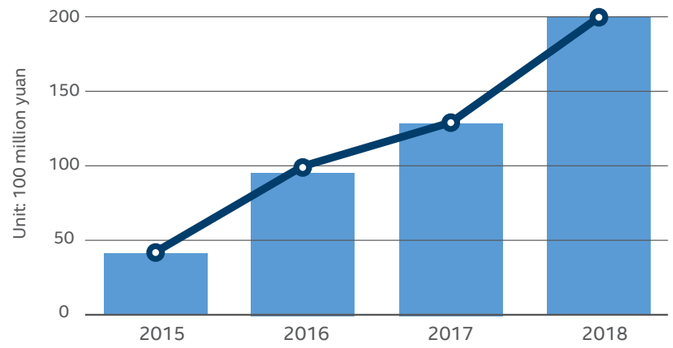


Figure 1-1-1 Market Size of China's Medical AI Industry

[1] **2018 Blue Paper on World AI Industry:** http://www.semi.org.cn/siip/pdf/20180920p2.pdf
[2] Outlook Industry Research Institute, **Analysis Report on the Market Outlook and Investment Strategy Planning of China's AI Industry from 2018 to 2023,** 2018: https://bg.qianzhan.com/report/detail/300/190314-389cc4a4.html
[3] Global Market Insights Report, April 2018: www.elecfans.com/rengongzhineng/592041.html
[4] Outlook Industry Research Institute, **Report on the Market Outlook and Investment Strategy Planning of China's Gene Sequencing Industry from 2018 to 2023,** 2018: https://bg.qianzhan.com/trends/detail/506/180411-e7daa2c4.html

of Internet + Healthcare, identifying the promotion of "Internet + AI" application services as an important measure to implement the "Healthy China" strategy, and stating that the government would focus on supporting the research and development of healthcare related AI technologies, medical robots, large medical equipment, and so on.

Application Trend of AI in Healthcare

AI has a wide range of applications in healthcare. From medical imaging, auxiliary diagnosis and disease prediction, to health management, drug R&D, chronic disease management and epidemic prevention and control, AI can play a key role, and has already served different functions in healthcare organizations at different levels and in different segments. Among them, the application of AI in medical imaging, auxiliary diagnosis and disease prediction is primarily implemented by hospitals and other healthcare organizations, and focuses on disease screening, as well as how to improve diagnostic accuracy. However, due to limitations of false negative results, doctors are still required to review all images to avoid missed diagnosis, so the effect of such application in reducing the workload of doctors is not significant.

In the future, the application of AI in healthcare organizations at different levels may become more diversified. That is to say, in primary hospitals or third-party physical examination centers, its application will focus on auxiliary screening, auxiliary diagnosis and chronic disease management; in 3A-class hospitals, its application will aim to improve the productivity of doctors; in the field of health management, AI will mainly provide support for the physical examination service paid by organizations and individuals; and in the field of drug R&D, AI application requires close cooperation between related technology companies, large pharmaceutical enterprises, and pharmaceutical research institutions.

Although AI is rapidly adopted in the healthcare industry, it still faces many challenges due to limitations of data, models and so forth.

- **Data volume.** The more complex the model and the more parameters, the more training samples are needed. However, in many complicated clinical scenarios, it's not easy to obtain a large amount of reliable data.
- **Data dimension.** Generally, the fewer the data dimensions, the worse the ability to describe the real world, but high-dimensional data processing has drawbacks such as low processing efficiency and high computational requirements.
- **Data quality.** Generally speaking, the degree of structuring and standardization of health data is low, and these data is often discrete and noisy. In under-equipped clinics and primary hospitals, there are still problems such as missing or wrong electronic medical records, decentralized data storage across multiple departments, and unreliable data.
- **Data silos.** Medical data is very sensitive due to its private nature, so healthcare organizations take the risk of medical data breach very seriously, but this also gives rise to the phenomenon of data silos between different healthcare organizations. And it is difficult for a single healthcare organization to gather enough high-quality training data for AI model training and learning.
- **Interpretability of model.** The deep learning model is a black box, which is not able to give a clear explanation on how the model comes to its conclusion, and the authenticity of its decision-making mechanism is yet to be verified.
- **Generality of model.** First, there exists model bias. For example, a model trained with white patients' data, may not work well for patients of other races. Second, the interoperability of model is very poor. In other words, it is difficult to build a deep learning model suitable for two different electronic medical record systems.
- **Robustness of model.** Even a properly trained image processing model may be adversely affected by the noise of the input image, and such noise is hard to be noticed by human. In addition, a minor data change may have significant impact on prediction results. For example, slight changes to experimental test values in patients' electronic medical record will greatly affect the prediction results of the model on hospital mortality.

In response to these challenges, several measures are taken by experts in both AI and healthcare to optimize the application environment and improve its effectiveness.

- **Collect large-scale and diversified health data.** Widely collect, standardize and integrate data from patients with different races, nationalities, languages and socioeconomic status;
- **Improve data quality.** Start with reliable, high-quality data input and then use tools to improve the quality of data collection, such as error correction, warnings about missing data.
- **Integrate into clinical workflow.** Integrate deep learning into the management of existing electronic medical record systems to improve clinician productivity and real-time data collection.
- **Build a high-dimensional learning model.** Introduce millions and even billions of rules to dramatically improve the accuracy of prediction and identification by using high-dimensional learning models.
- **Legislation and standardization.** To address information security issues such as computer hackers tampering with data and thus affecting the results of deep learning models, regulations are needed to protect analytical models.

At the same time, in order to promote safer interaction, transmission and aggregation of multi-source medical data, and to solve the problem of insufficient high-quality training data caused by data silos, experts from all fields are actively exploring the introduction of more secure data collaboration methods (such as the federated learning method) and better training architectures for AI model, so as to build a safe and reliable multi-source data collaboration solution with more high-quality data while reducing the risk of privacy breach, thereby enhancing the performance of medical AI applications, and enabling AI to serve the healthcare industry in a more efficient and secure way.

Application Scenarios of AI in Healthcare

Healthcare is one of the most promising applications for AI, which is a common belief in the industry. As its application continuously extends, AI will play an important role in the following application scenarios in healthcare:

- **Chronic disease management and disease monitoring.** Evaluate the risks of (potential) chronic diseases based on signs and symptoms of patients, so as to greatly reduce the medical expenses of patients through early intervention.
- **Clinical prediction and analysis.** For example, evaluate the risk of nosocomial infection (such as septicemia) based on electronic medical record data, predict the readmission rate of patients according to the operation model, and develop bundled services according to the financial model.
- **Chronic disease management.** Use data collection methods (e.g. Internet of Things) to build an AI-based chronic disease assessment and screening model, thereby improving the prediction and early diagnosis of chronic diseases.
- **Medical record search and quality control.** Accurately extract key information in texts of medical records and recognize medical entities to enable flexible and full electronic medical record search.
- **Virtual reality assistant.** With the aid of virtual reality, educate patients, help patients clearly understand the cause of their diseases and make doctor-patient communication more effective.
- **Intelligent guidance.** With the aid of multiple interactive modes such as voice and touch screen, provide better in-hospital navigation and guidance to improve triage, reception, health consultation and health education services.
- **Image-assisted diagnosis.** Facilitate radiologists quickly screen out normal images to improve the doctor's productivity; improve the accuracy of image analysis, shorten the reporting time of diagnostic results, and improve the diagnostic capability of the medical system.
- **Pathological analysis.** For example, efficiently and accurately detect and classify cancer cells, and accurately delineate targets for cancer radiotherapy.
- **Genomic analysis.** Dramatically reduce the cost of gene sequencing and rapidly and accurately analyze genomic data on a large scale to support the diagnosis and treatment of diseases such as cancer.
- **Drug Research.** Speed up drug research and development and reduce costs.
- **Prevention and control of epidemics.** AI can be used to model infection pathways, simulate networks of potential infections, and find possible transmission pathways to aid in precise prevention and control while accelerating the development of vaccines and drugs.

In the next chapter, "Implementation", we will detail our cooperation with Neusoft, Siemens, Accurad, 4Paradigm, Huiyi Huiying, Vistel, Winning Health and KFBIO in medical AI implementation projects, including project background, implementation process, experience and achievements, and recommend corresponding software and hardware configurations in various application scenarios.

Implementation

OpenVINO™ Improves AI Inference Efficiency for Medical Images

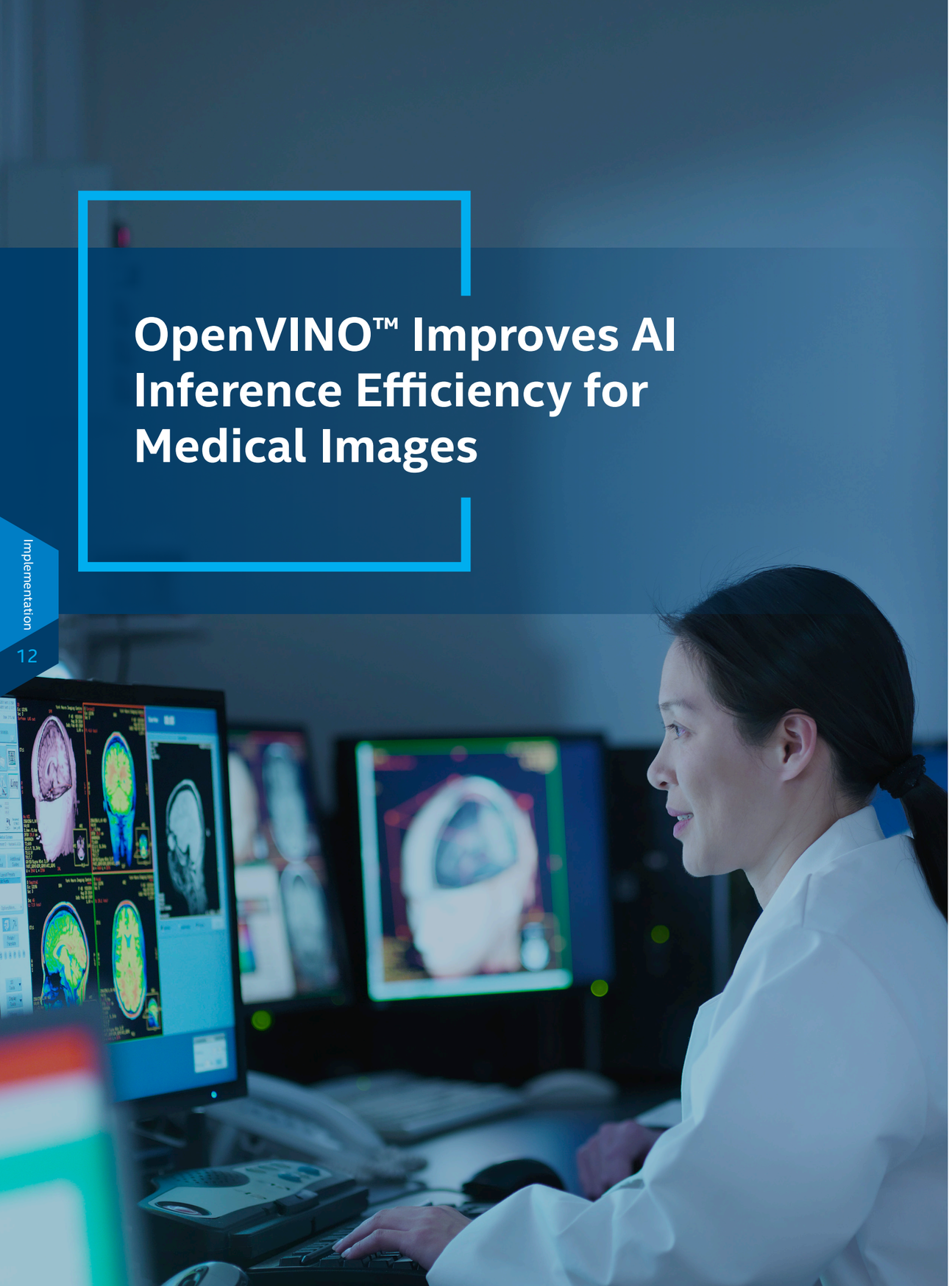


Image Segmentation in Medical Imaging

Traditional Medical Image Segmentation Method

Image segmentation in computer vision[5] refers to the segmentation of an image into multiple regions by natural boundaries in the image, such as object contours and lines, with an aim to simplify or change the representation of the image, making it easier to interpret and analyze. When computerizing, this process is usually implemented by labeling each pixel in the image to ensure that pixels with the same label have some common visual characteristics, such as color, brightness and texture, and finally, a certain area obtained from the above measurement or calculation will have similar pixel characteristics, while its adjacent areas are quite different in pixel characteristics.

As an important branch of computer vision technology, image segmentation has been widely used in many industries and fields such as medical image processing, face recognition, industrial robots, intelligent transportation, fingerprint recognition and satellite image positioning. In the field of medical image processing, image segmentation has proved its value in many areas such as tumor and other pathological site targeting, tissue volume measurement, anatomical research, computer-assisted surgery, treatment plan development and auxiliary clinical diagnosis.

Typical traditional image segmentation methods mainly include:

- **Clustering-based method:** The clustering method is based on K-means algorithm, which divides the image into K clusters through iterations. In this algorithm, there are similar distance deviations between the pixels in the segmented image and the clustering center, and the distance deviation is usually represented by indicators such as color, brightness, texture and location. The algorithm has good convergence.
- **Threshold-based method:** This method calculates one or more gray scale thresholds of an image and compares the gray scale of each pixel with the thresholds, and then performs clustering based on such comparison.
- **Edge-based method:** This method segments the image according to the sudden change in gray scale, color, texture and other characteristics of natural edges in the image. Generally speaking, the edge-based segmentation method relies on the detection of edge gray scale, and takes the edge with step-like change in gray scale as an image edge.
- **Region-based method:** This method segments the image according to the similarity of image, and expands the set of pixel points if adjacent pixel points have similar gray scale, color, texture, etc.

Deep Learning-based Image Segmentation Method

With the rapid development of AI technologies in recent years, especially in the field of image processing, the AI-based image recognition and image processing applications have been used in many scenarios, and their capabilities in analyzing and recognizing various medical images have surpassed those of human beings. Currently, several deep learning-based network models similar to Convolutional Neural Network (CNN) are widely used in AI-based image segmentation, which include Fully Convolutional Network (FCN), U-Net and V-Net.

■ FCN

A typical use of CNN is to classify tasks. When used in image processing, its output is a single category label. In biomedical image segmentation, the desired output should include location, i.e. the category label should be assigned to each pixel. As an upgraded and expanded version of Convolutional Neural Network, FCN[6] follows a network structure consisting of encoding and decoding and cascades the convolutional layer and the pooling layer, as shown in Figure 2-1-1. The convolutional layer and the max-pooling layer can effectively reduce the spatial dimension of the original image. At the same time, FCN uses AlexNet as the network encoder, and up-samples the feature map output from the last convolutional layer of the encoder by performing multiple transposed convolution until the feature map restores to the resolution of the input image. Thus, pixel-level image segmentation can be successful.

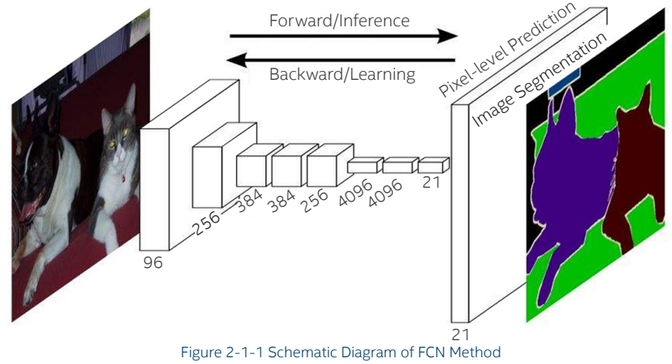


Figure 2-1-1 Schematic Diagram of FCN Method

[5] For the description of image segmentation, please refer to: Linda G. Shapiro and George C. Stockman (2001): "Computer Vision", pp 279-325, New Jersey, Prentice-Hall, ISBN 0-13-030796-3

[6] Relevant technical description of FCN is quoted from **Fully Convolutional Networks for Semantic Segmentation** written by Jonlong, Shelhamer and Trevor from UC Berkeley: https://people.eecs.berkeley.edu/~jonlong/long_shelhamer_fcn.pdf

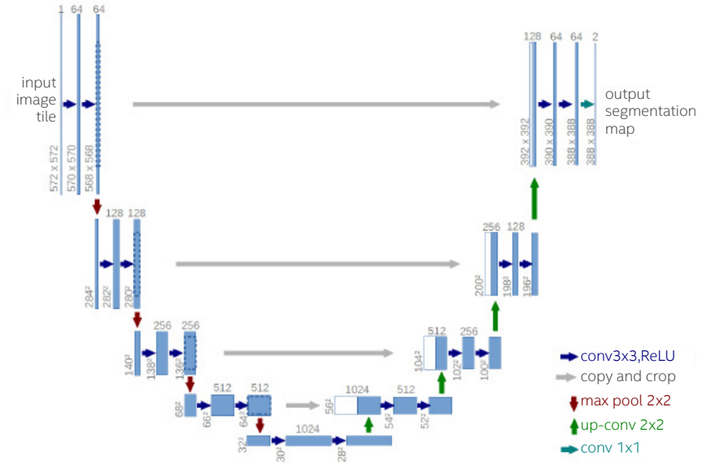


Figure 2-1-2 U-Net Topology

■ U-Net

As an improved version of FCN, U-Net has a distinct U-shaped structure, and its topology diagram is shown in Figure 2-1-2. It performs 4 up-samplings on each encoder, which allows for a more precise restoration of segmented image edge. At the same time, on the same stage, U-Net uses skip connection, instead of direct supervision and loss backpropagation on high-level semantic features, which ensures that the final restored feature map not only incorporates more low-level features, but also allows for features of different scales to be fused, thus enabling multi-scale prediction and deep supervision. In addition, at the rear section, U-Net adds a network similar to the front section, forming a U-shaped structure. The pooling operator is replaced by the up-sampling operator, thus increasing the resolution of the output. At the same time, in order to locate, the model combines the high-resolution feature of the shrinkage path with the up-sampled output. The continuous convolutional layer can use the **ReLu** activation function to down-sample the original image to obtain more precise output.

Medical imaging also has its unique characteristics in practical application. We can see that the typical chest radiograph is imaged for the chest, while eye fundus is imaged for the eye fundus OCT, both of which are imaging for a designated organ, not the whole body. However, the structure of the organ itself is relatively fixed and its semantic information is not particularly rich. Therefore, high-level semantic information and low-level features are very important, and U-Net's U-shaped structure and skip connection can play a greater role in this scenario. In recent years, the results of U-Net in medical image segmentation have been fully proved in many implementations.

■ V-Net

V-Net can be regarded as a 3D version of U-Net, which has a similar topology to U-Net and is suitable for segmentation of medical image with 3D structure, as shown in Figure 2-1-3. V-Net can perform end-to-end image semantic segmentation based on 3D images, and improve the network by using trick similar to residual learning.

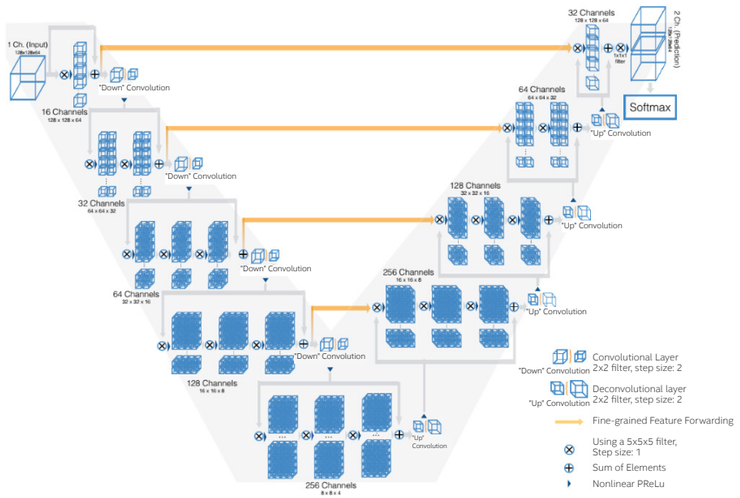


Figure 2-1-3 Concept of V-Net Topology

Recommended Hardware and Software Configuration

When building a deep learning-based image segmentation solution in healthcare industry, please refer to the following hardware and software configurations based on Intel® platform.

Name	Specification
Processor	Intel® Xeon® Gold 6240 processor or higher
Hyper Threading	On
Tutbo Boost	On
Memory	16GB DDR4 2666MHz* 12 and above
Storage	Intel® SSD D5-P4320 Series and above
Operating System	CentOS Linux 7.6 or latest version
Linux Kernel	3.10.0 or latest version
Compiler	GCC 4.8.5 or latest version
Python Version	Python 3.6 or latest version
TensorFlow Version	R1.13.1 or latest version
OpenVINO™ Toolkit	2019 R1 or latest version
Keras Version	2.1.3 or latest version

Optimizations for U-Net

Optimizations Based on Intel® Architecture

When the traditional CNN image segmentation method is applied to medical images, the following difficulties often exist:

- CNN is usually used for classification tasks, while biomedical images are more commonly processed in segmentation and localization tasks.
- CNN requires a large amount of training data, but it's very difficult to obtain a large amount of medical images.

In the past, when dealing with the above difficulties, sliding window method was usually employed, which means that, for each pixel to be classified, a part of its neighborhood will be extracted as input. The advantages of this method are as follows: first, this method can perform localization while sliding window; second, every operation will extract a neighborhood around a pixel, which can greatly increase the amount of training data. However, this method also has two disadvantages: first, there is a large overlap between the blocks extracted by sliding window, which will slow down training and inference; second, the network needs to make a trade-off between local accuracy and context acquisition, because if the block extracted by sliding window is too large, more pooling layers will be required and the localization accuracy will decrease, while if the block extracted is too small, the network can only see a small part of the context.

A series of optimizations based on Intel® platform can facilitate users solve the above problems from a different perspective. These optimizations include: adjusting the number of processor cores, and introducing Non-Uniform Memory Access Architecture (NUMA) technology and Intel® Math Kernel Library for Deep Neural Networks (Intel® MKL-DNN), thereby providing multi-level optimizations for U-Net. The optimization steps are as follows:

■ Setup Environmental Variables

First of all, users need to set up environment variables by using the following commands: clear system cache, set the processor to the performance priority mode, i.e. running at the highest frequency, and turn on the turbo boost of the processor.

```
echo 1 > /proc/sys/vm/compact_memory
echo 3 > /proc/sys/vm/drop_caches
echo 100 > /sys/devices/system/cpu/intel_pstate/min_perf_pct
echo 0 > /sys/devices/system/cpu/intel_pstate/no_turbo
echo 0 > /proc/sys/kernel/numa_balancing
cpupower frequency –set –g performance

export KMP_BLOCKTIME=1
export KMP_AFFINITY=granularity=fine,verbose,compact,1,0
export KMP_OMP_NUM_THREADS=20
```

- Set **KMP_BLOCKTIME** to 1 to specify the time, in milliseconds, that a thread should wait, after completing the execution of the current task, before sleeping.
- Set **KMP_AFFINITY** to Compact, which means that in this mode, thread binding takes precedence according to the computing requirements of the computing core, i.e. binding the same core first, and then sequentially binding the next core on the same processor. This kind of binding is suitable for the computation with data exchange or common data between threads. Its advantage is that it can make full use of the characteristics of multi-level cache;
- Set **OMP_NUM_THREADS** to 20 to specify the number of parallel execution threads to the required number of physical cores.

■ Add Thread Control to the Test Code

```
config = tf.ConfigProto()
config.allow_soft_placement = True
config.intra_op_parallelism_threads = FLAGS.num_intra_threads
config.inter_op_parallelism_threads = FLAGS.num_inter_threads
```

As shown in the above setup, during initialization of **tf.ConfigProto()**, we can also control the number of threads that each operator computes in parallel by setting the **intra_op_parallelism_threads** parameter and the **inter_op_parallelism_threads** parameter. The difference is:

- **intra_op_parallelism_threads controls operator internal parallelism**. When there is a single operator, and internal parallelism is possible, such as matrix multiplication, **reduce_sum** and so on, you can set the **intra_op_parallelism_threads** parameter to implement parallelism. "intra" means internal.

- **inter_op_parallelism_threads** controls parallel computation among multiple operators. When there are multiple operators which are independent of each other and there is no direct Path connection between operators, TensorFlow will try to compute them in parallel using a pool of threads, whose number is controlled by the **inter_op_parallelism_threads** parameter.

Normally, **intra_op_parallelism_threads** is set to the number of physical cores of a single processor, while **inter_op_parallelism_threads** is set to 1 or 2.

■ Utilize the Characteristics of NUMA to Control the Use of Processor Computing Resources

The servers used in a data center are usually equipped with two or more processors, and most of data centers use NUMA technology to run many servers as if they were a single system. A processor can access its own local memory faster than non-local memory. In order to obtain better computing performance in such a system, it needs to be controlled by some specific instructions. numactl is a technical mechanism for controlling processes and shared storage and is widely used for controlling computing resources in Linux. The specific usage is as follows:

Figure 2-1-4 Utilize the Characteristics of NUMA to Control the Use of Processor Computing Resources

```
1. numactl –C 0-19,40-59 –m 0 python3 test,py
```

The above command indicates that when executing, **test.py** only uses the 0-19 and 40-59 cores in processor #CPU0 as well as only uses the near-end memory corresponding to processor #CPU0.

■ Use TensorFlow Optimized for Intel® MKL-DNN

In order to enable users to perform efficient AI computing on general-purpose processor platforms, Intel has made a large number of optimizations for many mainstream open source deep learning frameworks, including TensorFlow, which is widely used in industry and academia.

Intel optimized TensorFlow by using a variety of primitives optimized by Intel® MKL-DNN. Intel® MKL-DNN was added from TensorFlow 1.2. In addition to significantly improving performance when training CNN-based models, compiling using Intel® MKL-DNN can also create binary files optimized for Intel® Advanced Vector Extensions (Intel® AVX), Intel® AVX 2, and Intel® AVX-512, resulting in an optimized file that is compatible with most modern (post-2011) processors.

References:

- https://www.tensorflow.org/guide/performance/overview?hl=en_us
- https://software.intel.com/content/www/us/en/develop/articles/tensorflow-optimizations-on-modern-intel-architecture.html

* For more technical details of Intel® MKL-DNN, please refer to relevant introduction in the Technologies section of this guide.

Test and Results of U-Net Optimized with Intel® Architecture

With the aid of above four optimizations, the performance of U-Net on Intel® processor platform has been significantly improved, and the test results are shown in the following figure[7]:

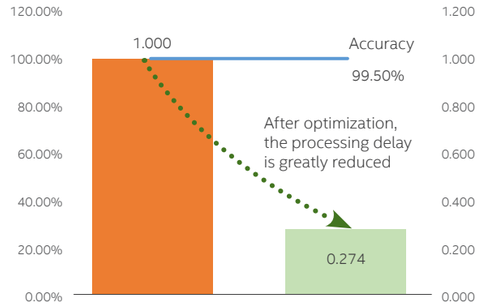


Figure 2-1-5 Performances with/without Intel® Optimizations

Further Optimizing U-Net with Intel® Distribution of OpenVINO™ Toolkit

In addition to the above achievements, in order to meet customers' requirements in actual application scenarios, Intel has further optimized the U-Net image segmentation method by using Intel® Distribution of OpenVINO™ toolkit (hereinafter referred to as the "OpenVINO™ toolkit"), with specific optimization steps as follows:

■ Convert the model

Since the original model is trained based on Keras and the generated model is in hdf5 format which cannot be directly used as the input of OpenVINO™ toolkit, so the hdf5 format is required to be converted by using the following command:

```
1. git clone https://github.com/amir-abdi/keras_to_tensorflow.git
2. cd keras_to_tensorflow
3. python3 keras_to_tensorflow.py --input_model=../unet/unet_membrane.hdf5 --output_model=../unet/full_unet.pb ##做模型转换
```

■ Convert the model into xml file and bin file by using mo.py included with OpenVINO™ toolkit

The command is as follows:

```
1. python3 /opt/intel/openvino/deployment_tools/model_optimizer/mo.py --framework tf --input_model full_unet.pb --data_type FP32 --output_dir ./ --input_shape [28,512,512,1]
```

■ Perform model inference by using Inference Engine

The command is as follows:

```
1. python3 segmentation_demo.py -m /home/worker/unet/full_unet.xml -i /home/worker/0.png -l /home/worker/openvino/intel64/Release/lib/libcpu_extension.so
```

In which, the inference code includes the following logic modules:

```
1. #Load input data ##数据数据预处理及导入，包括数据格式统一（医学图像 dcm 格式转化为 jpg）； 执行图像缩放， 多通道扩增，归一化处理等操作；
2. #Loading model to the plugin net = IENetwork.from_ir(model=model_xml, weights=model_bin) ##其中 xml 文件为网络结构， bin 文件为权重参数）
3. input_blob = next(iter(net.inputs))##确定模型的输入
4. out_blob = next(iter(net.outputs)) ##确定模型的输出
5. exec_net = plugin.load(network=net) ##模型导入
6. # Start sync inference
7. res = exec_net.infer(inputs={input_blob: images}) ##数据推理过程
8. # Processing output blob
9. res = res[out_blob] ##提取推理结果
10. #Visualization result
```

Results of Optimization with OpenVINO™ toolkit

As shown in Figure 2-1-6, the leftmost column is the original image of brain CT, the middle column is the image segmentation result before optimization, and the rightmost column is the image segmentation result optimized by using OpenVINO™ toolkit. It can be seen that, the image segmentation result generated and optimized by using OpenVINO™ toolkit is basically the same as that of the unoptimized image segmentation result in terms of accuracy rate, but its inference speed is much higher than that of the unoptimized image segmentation result[8].

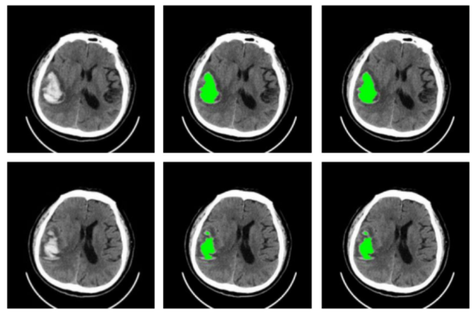


Figure 2-1-6 Results of U-Net Optimization with OpenVINO™ Toolkit

* For more technical details of OpenVino™ toolkit, please refer to relevant content in the Technologies section of this guide.

Dense U-Net Built on 2nd Generation Intel® Xeon® Scalable Processor

Intel® Deep Learning Boost (Intel DL Boost) Technology

The 2nd Generation Intel® Xeon® Scalable Processor not only improves computing performance with optimized micro-architecture, more cores and faster memory channels, but also provides more comprehensive acceleration capability for AI applications. In particular, it adds support for INT8 to its integrated Intel® DL Boost (VNNI instruction set), providing users with efficient INT8 deep learning and inference acceleration capability, which will effectively improve the execution efficiency of U-Net image segmentation method.

Intel® DL Boost supports 8-bit or 16-bit low-precision numerical multiplication through VNNI instruction set, which is especially important for deep learning computing that require a large number of matrix multiplications. The introduction of Intel® DL Boost can reduce user's requirement for system memory by up to 75%[9] when performing INT8 inference. The reduction of required memory and bandwidth also speeds up the low-precision numerical calculation, thus greatly improving the overall performance of the system.

Compared with the previous FP32 model, the INT8 model has smaller numerical precision and dynamic range. Therefore, when using the INT8 inference in deep learning such as image segmentation, users should pay more attention to resolve information loss during calculation and execution. Generally speaking, the INT8 inference function can generate a INT8 model to be inferred by using quantitative correction, thereby realizing the goal of converting FP32 into INT8 and minimizing information loss.

Taking the image analysis application as an example, the conversion from high-precision numerical value to low-precision data is actually a process of calculation and reduction. In other words, how to determine the range of reduction is the key to minimize information loss. In the process of mapping FP32 to INT8, the parameters of mapping reduction are determined by performing correction according to the data set. After determining the parameters, the platform will analyze the graph and perform quantization/inverse quantization and other operations according to the supported INT8 operation list. The quantization operation is used to quantize FP32 to S8 (signed INT8) or U8 (unsigned INT8), while the inverse quantization operation performs an inverse operation.

Conversion from FP32 model to INT8 model with OpenVINO™ toolkit

Generally, the model trained by neural network is of single-precision floating-point precision, namely FP32. Users can directly deploy such a model in actual application scenarios and obtain low precision models by

[7] The test configuration is as follows: Processor: 2S Intel® Xeon® Gold 6148 Processor, 2.40GHz; Cores/Threads: 20/40; Memory: 16 GB DDR4 2666 MHz * 12; Hard Disk: Intel® SSD SC2BB480G7; BIOS: SE5C620.86B.02.01.0008.031920191559; Operating System: CentOS Linux 7.6; Linux Kernel: 3.10.0-957.21.3.EL7.x86 _ 64; gcc Version: 7.2; Python Version: Python 3.6; TensorFlow Version: R1.13.1.

[8] The validation test configuration is as follows: Processor: 2S Intel® Xeon® Gold 6148 Processor, 2.40GHz; Cores/Threads: 20/40; Memory: 16GB DDR4 2666MHz * 12; Hard Disk: Intel® SSD SC2BB480G7; BIOS: SE5C620.86B.02.01.0008.031920191559; Operating System: CentOS Linux 7.6; Linux Kernel: 3.10.0-957.21.3.el7.x86_64; gcc version: 4.8.5; Python version: Python 3.6; OpenVINO™ Toolkit: 2019 R1; Keras: 2.1.3.

[9] Data Source: https://software.intel.com/en-us/articles/lower-numerical-precision-deep-learning-inference-and-training

using quantization technology. For example, INT8 model can provide more efficient model inference while ensuring model precision, with a loss of model precision normally less than 1%.

OpenVINO™ toolkit provides conversion from FP32 model to INT8 model starting from 2018 R4 version, and supports Intel® DL Boost integrated with the 2nd Generation Intel® Xeon® Scalable Processor starting from 2019 R1 version.

The basic work and deployment process for the Model Optimizer in the OpenVINO™ toolkit is as follows: the OpenVINO™ toolkit will first convert and optimize the trained Open Neural Network Exchange (ONNX) model to generate xml files and bin files in FP32 format in which the optimization step includes node fusion, batch normalization removal, constant folding, and so on. Then, the FP32 files will be converted into INT8 xml files and bin files by the conversion tool included with OpenVINO™ toolkit. During the conversion, a small verification data set will be used, and the statistical data in the conversion and quantization process will be stored in order to ensure that accuracy will not be affected during the subsequent inference. The above-mentioned conversion process is run offline. That means, only one conversion is required, which is detailed in Figure 2-1-7:

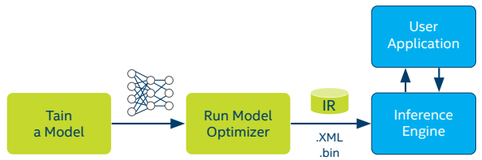


Figure 2-1-7 Conversion of FP32 Model to INT8 Model with OpenVINO™ Toolkit[10]

The performance of the preliminary model obtained after the above conversion is shown in the following figure:

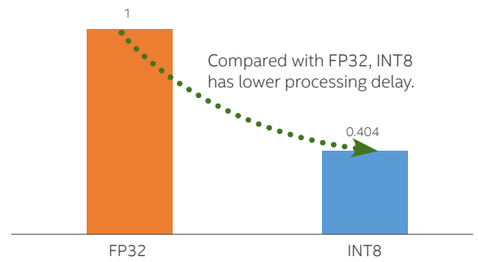


Figure 2-1-8 Comparison of Delay Performance between FP32 and INT8

Through the performance analysis of the two models, it can be seen that Reorder Ops takes up a large amount of execution time in the FP32 model. In the INT8 model, since Resample Ops only supports FP32 operations, the Concat Ops takes too long to execute, while Convolution Ops, which originally accounts for the highest proportion, takes up less time in the whole model operation. Therefore, it needs to be further optimized.

As shown in Figure 2-1-9, after optimization, the delay of the model has been greatly reduced.

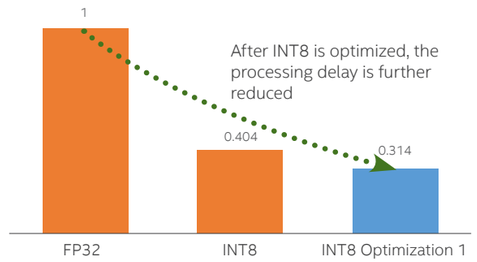


Figure 2-1-9 Comparison of Delay Performance of the Optimized INT8 Model

Now, when analyzing the INT8 model layer by layer, it can be seen that there has been a significant improvement compared with the previous model. However, in the optimized model, the execution time consumed by Concat Ops is still relatively long. In order to further improve the throughput of the model, Concat Ops needs to be specifically optimized and customized instead of using primitives in Intel® MKL-DNN. The detailed code is as follows:

```
1.  for (size_t i = 0; i < num_src; i++) {
2.      const MKLDNNMemory& src_mem = getParentEdgeAt(i)->getMemory();
3.      channels.push_back(src_num.GetDims()[1]);
4.      src_ptrs.push_back(reinterpret_cast<const uint8_t*>(src_mem.GetData()));
5.      dst_ptrs.push_back(dst_ptr + channels_size);
6.      channels_size += src_mem.GetDims()[1];
7.  }
8.
9.  parallel_for(iter_count, [&](int i){
10.     for (int j = 0; j < src_ptrs.size(); j++) {
11.         memcpy(dst_ptrs[j] + (i * channels_size), src_ptrs[j] + i * channels[j], channels[j]);
12.     }
13. });
```

The main purpose of the above optimization is to perform batch copying of data to specified locations in parallel. After this optimization, the model performance has been further improved. Now, the execution time of the model basically meets our expectation, and the final optimization results are shown in Figure 2-1-10:

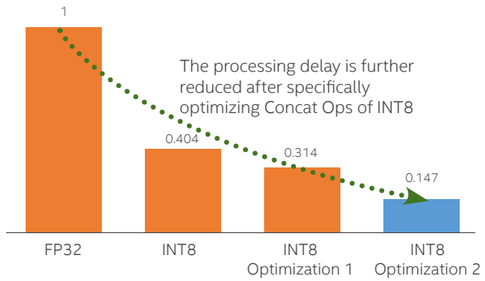


Figure 2-1-10 Comparison of Delay Performance of INT8 Model after Further Optimization

From the performance analysis, it can be seen that the primitive with the highest model execution time becomes a convolution operation, which is exactly the expected effect of Dense U-Net model in this example.

[10] This figure comes from https://docs.openvinotoolkit.org/latest/_docs_MO_DG_Deep_Learning_Model_Optimizer_DevGuide.html

Use Cases

Neusoft eStroke Thrombolysis and Embolectomy Imaging Platform

■ Background

Stroke has always been a major "killer" endangering public health. It is estimated that there are about 2 million new stroke patients in China every year in which nearly 50% of them are under 65 years old. This indicates that the stroke patients are becoming younger, and are still rising at a rate of 13% every year, with the recurrent rate reaching 17.7%[11], which has brought heavy burden to patients and society. Thrombolysis and thrombectomy are the preferred and effective treatment methods for stroke, but they rely on rapid and accurate interpretation of brain medical images.

The critical time for stroke treatment is only 30 minutes and there is basically no time for referral, so the primary district and county hospitals often play a key role in treating stroke. However, due to the insufficient technical capabilities of primary hospitals, the percentage of thrombolysis and thromboembolism treatments is low. In addition, because of varying interpretation abilities of doctors and shortage of professional imaging doctors, imaging experts in central hospitals are also overwhelmed. These all lead to the lack of effective imaging guidance for thrombolysis and thromboembolism, the inability to effectively identify salvageable tissues, and the elapse of effective time for rescuing patients.

In order to address this challenge, the healthcare industry needs a tool that can quickly and accurately analyze related medical images even if the interpretation abilities of doctors in the primary hospital is insufficient. Now, the deep learning-based medical image interpretation has been gradually adopted by healthcare organizations to cope with the above problems. Neusoft Intelligent Medical Research Institute, Shenyang Neusoft Medical System Co., Ltd. (hereinafter referred to as "Neusoft") and many partners have jointly developed a high-quality eStroke thrombolysis and thrombectomy imaging platform, which can provide more accurate guidance for intravenous thrombolysis and arterial thrombectomy when treating acute stroke.

■ Solution and Results

eStroke thrombolysis and thrombectomy imaging platform is a cloud service platform that provides quantitative evaluation of ischemic stroke penumbra, cerebral microbleeds and collateral circulation as well as accurately assesses multimodal imaging in thrombolysis and thrombectomy with the following advantages:

- **Support multimodal imaging equipment,** including 16-row and above multislice Computed Tomography (CT), 1.5T and above Magnetic Resonance Imaging (MRI);
- **Achieve the automation of the whole process,** from the beginning of scanning sequence of hospital equipment, to the image post-processing and analysis, and to the output of the image diagnosis report, without manual intervention;
- **Access the Internet-based medical diagnosis and treatment technology application and research platform and other external diagnosis and treatment systems,** to support the development, research and engineering of cardiovascular and cerebrovascular diseases in remote first aid, mobile first aid, intelligent early warning and intervention for high-risk groups, joint treatment, virtual surgery, etc.

Based on eStroke thrombolysis and thrombectomy imaging platform, Neusoft and Intel jointly carry out image segmentation processing on stroke images in the platform by using the U-Net model, calculate various perfusion imaging parameters including CBF, CBV, MTT and TMAX (respectively corresponding to Cerebral Blood Flow, Cerebral Blood Volume, Mean Transit Time and Time To Peak of residual function), and rely on the symmetry of left and right cerebral circulation, as shown in Figure 2-1-11, to further infer the regions of ischemic penumbra and infarction core for medical diagnosis.

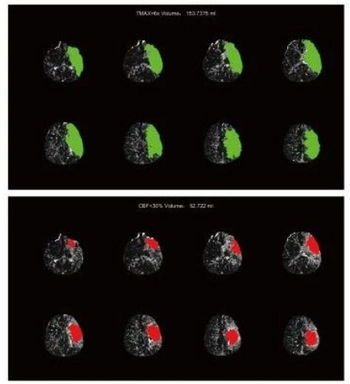


Figure 2-1-11 Calculate the Regions of Ischemic Penumbra and Infarction Core According to TMAX & CBF Abnormal Regions

The solution is optimized by using Intel® Optimization for TensorFlow (based on Intel® MKL-DNN optimizations) and OpenVINO™ toolkit, which enables the deep learning inference with U-Net model to ensure accuracy while reducing inference time greatly. This solution is undoubtedly of great significance for the time-sensitive stroke diagnosis and treatment. As shown in Figure 2-1-12, with basically the same inference accuracy, the solutions optimized with one of the above two tools reduce the inference delay by 72.6% and 85.4% compared to the unoptimized solution respectively[12].

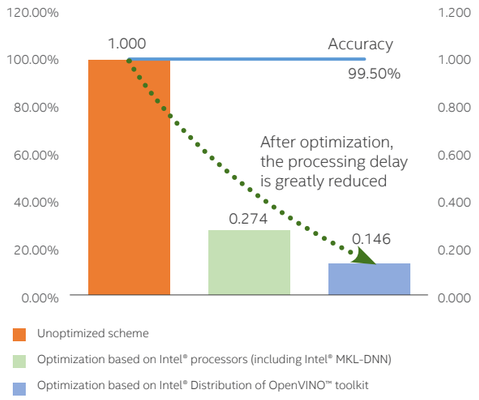


Figure 2-1-12 Performance Comparison of Neusoft U-Net Image Segmentation Solutions

[11] Data is quoted from **"Anhui Province Stroke Classification & Diagnosis and Treatment Guidelines (2015 Edition)"**

[12] The data is obtained in the following test configuration: Processor: 2S Intel® Xeon® Gold 6148 Processor, 2.40GHz; Cores/Threads: 20/40; Memory: 16 GB DDR4 2666 MHz * 12; Hard Disk: Intel® SSD SC2BB480G7; BIOS: SE5C620.86B.02.01.0008.031920191559; Operating System: CentOS Linux 7.6; Linux Kernel: 3.10.0-957.21.3.el7.x86_64; gcc Version: 7.2 (TensorFlow) & 4.8.5 (OpenVINO); Python Version: Python 3.6; TensorFlow Version: R1.13.1; OpenVINO™ Toolkit: 2019 R1; Keras: 2.1.3.

Siemens Healthineers Uses Intel® DL Boost to Advance AI Application in Cardiovascular Disease Treatment

■ Background and Challenges

Cardiovascular diseases have always been a major disease threatening people's health and life. According to statistics, about 18 million people lose their lives each year due to cardiovascular diseases[13]. The quantitative measurement of Cardiac Magnetic Resonance (CMR) images by using Magnetic Resonance Imaging (MRI) have always been the most efficient way for evaluating cardiac function, ventricular volume and myocardial tissue status. Traditionally, cardiovascular experts rely on their experience to interpret MRI images, which is not only time-consuming and laborious, but also vulnerable to subjective judgment and prone to error, resulting in missed diagnosis and misdiagnosis.

At present, Siemens Healthineers is carrying out a series of research on innovative medical AI applications and integrating the results into practical applications of cardiology and radiological image analysis. However, there are still a series of challenges to address when applying these AI capabilities.

First of all, AI application should not cause delay to clinical diagnosis and treatment. Instead, it should keep synchronization with data generated by various medical instruments and equipment and achieve high throughput and low delay for AI inference, so as to enable AI-based medical system to serve more patients. Secondly, AI application should be integrated with clinical diagnosis and treatment process as much as possible in order to save time and improve the consistency and accuracy between measurement and diagnosis.

For this reason, Siemens Healthineers and Intel employ a general processor platform to carry out MRI images interpretation and measurement and implement efficient AI inference. Specifically, they adopt a deep learning approach to perform AI-based interpretation on cardiovascular medical images from MRI, and utilize the optimizations provided by the brand-new 2nd Generation Intel® Xeon® Scalable Processor Platform and OpenVINO™ toolkit, thereby greatly improving the inference speed and providing strong support for clinical diagnosis and treatment.

■ Introduction to the solution and effects of its implementation

In this case, Siemens Healthineers cooperates with Intel to optimize the cardiac cavity detection and quantification model built on the brand-new 2nd Generation Intel® Xeon® Scalable Processor. This AI model is based on Dense U-Net, which can perform semantic segmentation on left and right ventricles and can be extended to all four chambers. The input of AI model is the stack of MRI images of the beating heart, and the output is the identified regions and structures of the heart, each of which is color coded. In this way, the original process of manual identification and labeling can be made smart and intelligent, thus accelerating the speed of image interpretation. The overall work flow is shown in Figure 2-1-13.

The 2nd Generation Intel® Xeon® Scalable Processor provides an efficient, flexible and scalable platform for inference of the AI model. In particular, its close collaboration with OpenVINO™ toolkit effectively accelerates deep learning inference for visual applications and improves the speed and accuracy of diagnostics and decision making that are critical in the diagnostic process. At the same time, the processor-integrated Intel® DL Boost provides a new Vector Neural Network Instruction (VNNI), which can further accelerate various computationally intensive operations in deep learning and provide more efficient AI inference for AI applications such as image classification, image segmentation, and object detection on Intel® processor platforms. Intel® DL Boost also provides better support for INT8, which enables it to convert FP32 training model into INT8, thereby greatly improving inference speed while maintaining accuracy.

In this case, a deep neural network (such as Dense U-Net) is used to identify the heart regions after training. The weight of the neural network is usually represented by floating point value (FP32), so the model is often trained and inferred with FP32 precision. However, INT8 can also improve inference speed with little loss of accuracy (usually < 0.5%, in this case < 0.001%)[14].

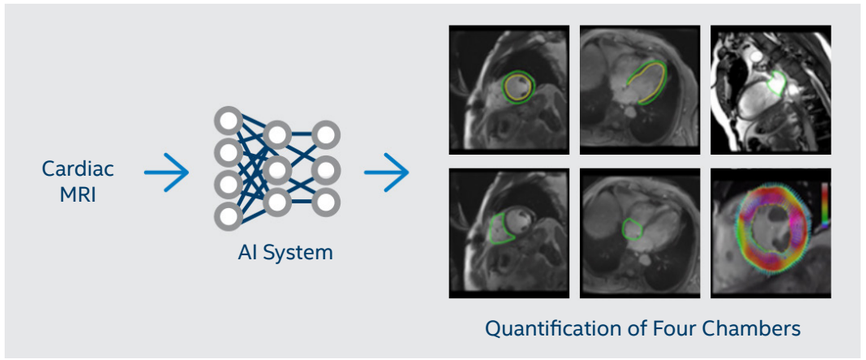


Figure 2-1-13 Siemens Healthineers and Intel Build AI Analysis Capability of Cardiac MRI

[13, 14] Data is quoted from Journal of the American College of Cardiology, 2017.

With the aid of Intel® DL Boost and the FP32 to INT8 conversion tool provided by the OpenVINO™ toolkit, Intel helps Siemens Healthineers speed up the inference operations while maintaining accuracy. Figure 2-1-14 shows the use of AI for heart image segmentation. The left figure shows that AI model segments various structures of the heart. The upper right figure shows the traditional ONNX output image without using the INT8 model, while the lower right figure shows the output image using the INT8 model, and it can be easily seen that both images have basically the same accuracy.

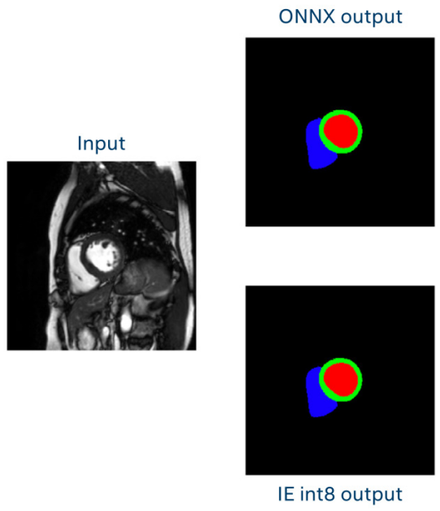


Figure 2-1-14 Comparison of Output Accuracy Before and After Using INT8 Model

From the perspective of inference speed, the AI analysis capabilities of cardiac MRI have been greatly enhanced by optimizations provided with the 2nd Generation Intel® Xeon® Scalable Processor, Intel® DL Boost and OpenVINO™ toolkit. On the one hand, the processing speed of cardiac MRI images has significantly improved, reaching 200 FPS (frames per second), which means that data from a complete cardiac MRI examination can be analyzed in less than 1 second and enables cardiac MRI to be used in near real-time. On the other hand, compared with the unoptimized solution, the performance of the optimized solution can be improved by 5.5 times[15] when quantifying and executing the model with almost no reduction in accuracy.

GE Healthcare Utilizes Intel Technologies and Products to Optimize Deep Learning Model and Improve Inference Performance in CT Imaging

■ Background

Computed Tomography (CT) scan is one of the most commonly used examination methods in modern medicine. It scans the human body with X-ray beams and obtains cross-sectional or 3D images of relevant parts, thereby discovering pathological changes of the human body. Although CT scan is extremely important, CT slice images are usually manually reviewed by experienced doctors, which is not only inefficient but also inherently subjective and may lead to misdiagnosis and missed diagnosis.

At present, GE Healthcare adopts the deep learning approach to classify and label CT slice images, which is more convenient for doctors to find small lesions and use them for research or clinical comparison. At the 2018 SPIE Optics of Medical Imaging Conference, GE Healthcare released a paper on AI-based structural classifiers, in which CT imaging experts from GE Healthcare used Python language, TensorFlow framework and Keras library to build and train a new AI model. Through in-depth technical cooperation with Intel, both parties are using products and technologies such as Intel® Xeon® processor and Intel® Deep Learning Deployment Toolkit (Intel® DLDT) to optimize their solutions for CT inference.

■ Solution and Results

Intel® DLDT is introduced to optimize the deep learning model and shows better inference performance on the Intel® Xeon® processor platform.

Included with OpenVINO™ toolkit, Intel® DLDT is an inference acceleration component that is specifically designed for deep learning models. With the aid of this tool, models which are trained for convergence can obtain higher data processing capability and lower data processing delay on various Intel® processor platforms. It can convert and optimize models trained by various mainstream open-source Deep Learning frameworks to generate bin files and xml files independent of the deep learning frameworks. The bin files are used to store the weight of the deep learning model, which is stored in binary form, while the xml files describe the network structure of the deep learning model. When parsing the model, both the bin files and the xml files will be used. This enables the representation file of the model to be independent of any deep learning frameworks and to be deployed more conveniently. At the same time, in the process of generating these two types of files, constant folding, Batch layer fusion, horizontal layer fusion, removal of invalid nodes and other model optimization operations will be carried out on the model.

[15] The data is obtained in the following test configuration: Processor: 2S Intel® Xeon® Platinum 8280 Processor, 2.70GHz; Cores/Threads: 28/56; HT: ON; Turbo: ON; Memory: 192GB DDR4 2933; Hard Disk: Intel® SSD SC2KG48; BIOS: SE5C620.86B.02.01.0008.031920191559; Operating System: CentOS Linux 7.6.1810; Linux Kernel: 4.19.5-1.el7.elrepo.x86_64; gcc Version: 4.8.5; OpenVINO™ Toolkit: 2019 R1; Workload: Dense U-Net.

As shown in Figure 2-1-15, Intel® DLDT can easily import GE Healthcare's models that are trained on frameworks such as TensorFlow.

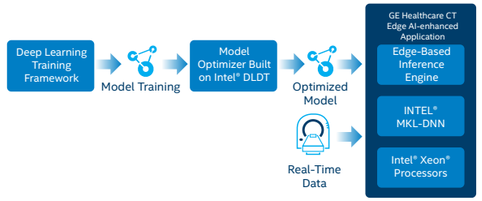


Figure 2-1-15 GE Healthcare CT Edge AI Enhancement Application with Intel® DLDT Deployed

After converting and optimizing the model with Intel® DLDT, the optimized model can be imported into GE Healthcare CT Edge AI Enhancement Application, which builds a powerful edge-based inference engine on Intel® Xeon® processor platform and Intel® MKL-DNN.

In order to verify the actual performance of this optimized solution, GE Healthcare and Intel carried out a series of performance tests with a data set consisting of 8,834 CT scan images. GE Healthcare hopes that after optimizing the model, the inference engine can process up to 100 images per second with less than 4 processor cores.

The test results show that on Intel® Xeon® processor E5-2650 v4 with only a single core enabled, the optimized model can increase the inference throughput to 14 times of that before optimization. At the same time, the multi-core performance of Intel® Xeon® processors has greatly improved the efficiency of GE Healthcare's medical inference engine. As shown in Figure 2-1-16, after using 4 processor cores, the number of images that the inference engine can process per second has increased to 596, nearly 6 times the original expectation.[16]

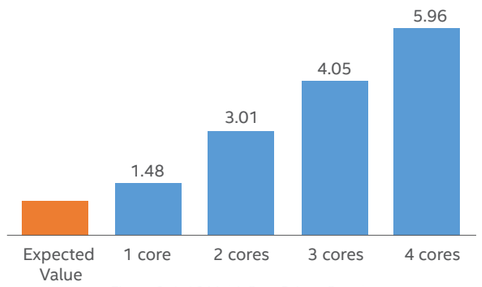


Figure 2-1-16 Multi-Core Brings Steady Improvement of Inference Performance

Huiyi Huiying Utilizes Intel Technologies to Build an Efficient AI-assisted Diagnosis and Treatment Platform

■ Background

Early screening and detection is an effective way to care for women's health and help keep them safe from breast cancer. Medical diagnosis can be assisted by ultrasound, X-ray scan, MRI, and other medical imaging techniques. As mentioned earlier, interpretation of images requires a doctor with extensive experience and an interdisciplinary knowledge base. However, even in some large hospitals, the number of doctors with these skills is insufficient, not to mention small community hospitals or healthcare organizations in outlying regions.

At the same time, although the growth in the number of medical images and the development of computer imaging technology have driven the emergence of computerized medical image analysis solutions, doctors tend to use them only for screening, classification and pre-determination prior to analysis and diagnosis, because the accuracy of traditional image diagnostic support system is not comparable to that of human interpretation. In addition, the lack of a unified data interoperability standard also leads to increased communication costs in the scenarios where a patient is treated by multiple doctors during the treatment period.

In order to facilitate healthcare organizations obtain a more efficient and intelligent auxiliary diagnosis and treatment platform, Huiyi Huiying, as a high-tech enterprise with the mission of empowering the tiered healthcare system and the precision medicine with artificial intelligence, cooperated with Intel to build the Dr. Turing AI, a deep learning-based auxiliary diagnosis and treatment solution, by introducing the OpenVINO™ toolkit and other advanced hardware and software products. Now, Dr. Turing AI has achieved satisfactory results in applications such as early breast cancer screening and diagnosis.

■ Solution and Results

As a new intelligent deep learning-based and image-assisted diagnosis solution, Dr. Turing AI can be used across the whole process of early breast cancer screening and diagnosis, and with its unified data interoperability, it also enables medical staff to improve the efficiency of image analysis, diagnosis, clinical detection support and disease management, showing several advantages:

- More accurate image analysis and multiple automatic labeling capabilities;
- Faster image-assisted analysis to improve the doctor's image review efficiency;
- Structured image report that complies with the standard of American College of Radiology (ACR);
- Patient information can be automatically updated in the breast image report and data system.

For higher image analysis accuracy, the solution can use various deep learning algorithm models, such as Inception V4, Inception ResNet V2 and other models, as needed. In some recent applications, as shown in Figure 2-1-17, the solution uses the RetinaNet object detection model with the ResNet50 convolutional network model as the backbone to implement model training and inference, where the ResNet50 convolutional network model is used to extract features and the subnetwork is used for classification and regression.

[16] The data is obtained in the following test configuration: Processor: Intel® Xeon® Processor E5-2650 v4, 2.20GHz; Cores/Threads: 12/24; HT: ON; Turbo: ON; Memory: 264GB; Hard Disk: 480GB; Operating System: CentOS Linux 7.4.1708; Linux Kernel: 3.10.0-693.el7.x86_64; gcc version: 4.8.5; Workload: Data set consisting of 8,834 CT scan images.

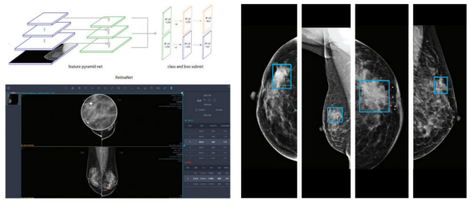


Figure 2-1-17 Solution Built on the RetinaNet Model

To further increase the speed of analysis, the new solution also introduces the OpenVINO™ toolkit to improve inference performance. On the one hand, the OpenVINO™ toolkit provides a series of built-in optimization tools and pre-trained models, which can be invoked by users to compress and accelerate the trained models, thereby improving the model inference efficiency; on the other hand, the solution can also use the OpenVINO™ toolkit to complete the conversion from FP32 to INT8 model, in exchange for a significant increase in inference speed with a controllable loss of model accuracy (taking the image classification for example, the accuracy loss of industry common model is less than 1%).

The Keras FP32 floating-point model with high accuracy is used for the training process, then in the inference process, the original model is converted to IR file by the Model Optimizer in the OpenVINO™ toolkit and input to the Inference Engine for inference, during which the built-in Calibration Tool is also used to quantize the FP32 model into the INT8 type to improve the inference speed.

As shown in Figure 2-1-18, inference on the FP32 model using the OpenVINO™ toolkit is 3.02 times faster than the original model, and after the INT8 conversion using the OpenVINO™ toolkit, the inference speed is improved by 8.24 times, while the accuracy loss is less than 0.17%[17].

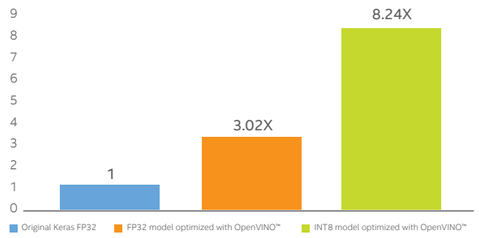


Figure 2-1-18 Improved Inference Efficiency with the OpenVINO™ Toolkit

As we all know, the more advanced the cancer is, the more medical resources are needed. Therefore, early detection and intervention of cancer can not only serve the patient better, but also save more medical resources to improve the health of the whole population. Now, the Dr. Turing AI, an Intel-powered auxiliary breast cancer diagnosis and treatment solution from Huiyi Huiying, has been deployed in a range of medical organizations, and has been praised by doctors and patients alike for not only enabling 8.24x faster image analysis[17] but also helping to reduce false positives and unnecessary biopsies of lumps and calcifications.

Winning Health Builds an Efficient Pulmonary Nodule Intelligent Auxiliary Diagnosis System based on Advanced Intel Products

■ Background

In the clinical diagnosis of lung cancer and other lung diseases, CT images of pulmonary nodules are not only an important basis for diagnosis, but also provide key information for the determination of treatment plans. Pulmonary nodules have common but very complex clinical features, their etiology is complex, and their clinical manifestation lacks specificity and is susceptible to the doctor's experience and subjective judgment. Therefore, the interpretation and analysis of images of pulmonary nodules need to be very detailed and accurate, which requires a high level of doctor diagnostic skills and medical resources.

Introducing AI into intelligent auxiliary diagnosis of pulmonary nodules can facilitate healthcare organizations effectively address this challenge. To this end, Winning Health Technology Group Corporation (hereinafter referred to as Winning Health), together with Intel and AMAX, has built a new intelligent deep learning-based auxiliary diagnosis system for pulmonary nodules. The system's intelligent auxiliary diagnosis model is interconnected with the Radiology Information System (RIS) and the Picture Archiving and Communication Systems (PACS), which can insert the quantitative imaging of pulmonary nodules into the RIS report and show the relationship between pulmonary nodules, surrounding tissues and blood vessels with the aid of intelligent 3D reconstruction.

For better deployment and operational performance, Winning Health selected the AMAX Deep Learning Appliance built on the 2nd Generation Intel® Xeon® Scalable processors and the OpenVINO™ toolkit as the infrastructure. The new processor not only has powerful general-purpose computing power, but also integrates innovative technologies such as Intel® AVX-512 and Intel® DL Boost, which provide a good balance between general-purpose computing power and parallel computing power, providing excellent performance for AI training. And the OpenVINO™ toolkit contains a large number of pre-trained models tuned and packaged by Intel for direct invocation by the user. In addition, the user can also utilize the OpenVINO™ model converter for numerical conversion in order to improve efficiency (see the section "Conversion from FP32 model to INT8 model with OpenVINO™ toolkit" on page 17 for more details).

As shown in Figure 2-1-19, test data in the tasks of segmentation, detection, and removal of false positives indicate that the OpenVINO™ toolkit can increase inference speed by a factor of 10-30[18].

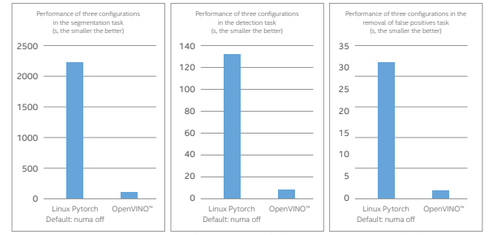


Figure 2-1-19 Performance of the Pulmonary Nodule Intelligent Auxiliary Diagnosis System in Different Tasks

[17] Internal test data from Huiyi Huiying: https://builders.intel.com/docs/aibuilders/huiying-medical-technology-optimizes-breast-cancer-early-screening-and-diagnosis-with-intel-ai-technologies.pdf, the test configuration is as follows: Processor: 2S Intel® Xeon® Platinum 8268 Processor, 2.90GHz; Cores/Threads: 24/48; Version of OpenVINO™ toolkit: Intel distribution 2019R2, the dataset contains 366 mammographic images provided by Huiyi Huiying, Image size 1280X640.

[18] The test configuration is as follows: 2S Intel® Xeon® Gold 6240 Processor, 18 cores/36 threads, HT Technology enabled; Total Memory: 384 GB (12 slots/32GB/2666MHz); Storage: Intel® SSDs D3-S4510; BIOS: SE5C620.86B.02.01.0010.010620200716 (ucode: 0x400002C), CentOS 8, Kernel: 5.6.4-1.el8.elrepo.x86_64; Deep Learning Framework: PyTorch; Compiler: gcc 7.3; MKL DNN Version: v0.20.5; Precision: FP32, Dataset: 357x4x3x96x512x512; Custom 3D U-Net; Configuration 1: Linux PyTorch (1.3.0) Default Numa OFF,1 instance; Configuration 2: Linux PyTorch (1.3.0) Optimized Numa ON, 36 instances; Configuration 3: OpenVINO, Version: 2019.3.376.

Vistel Launches an Intelligent Remote Medical Image Review Solution with Intel Technologies

When using traditional medical information systems, hospitals temporarily store the collected medical images into ImageHub, upload them to cloud servers for analysis and processing, and then return the results to hospitals' application software to help doctors diagnose diseases. As shown in Figures 2-1-20, the speed of feedback of results in this process may be limited by network factors as well as the speed of inference, which affects the efficiency of the treatment.

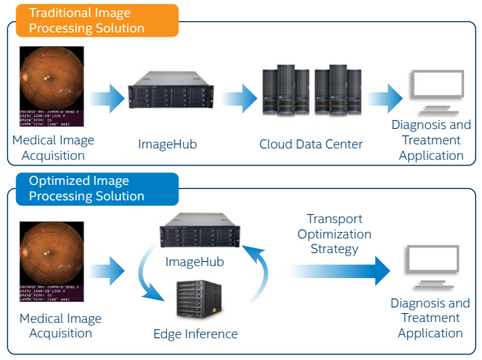


Figure 2-1-20 Comparison of Legacy and New Solutions for Intelligent Remote Image Review

To solve this problem, Beijing Vistel Technology Co., Ltd. (hereinafter referred to as "Vistel") deploys the Intel® Movidius™ Neural Compute Stick + OpenVINO™ toolkit on the edge side close to the onsite medical services through architecture optimization, as shown in Figure 2-1-20, to provide the front-line AI inference capability, allowing the solution to perform the compression, acceleration and inference process on the edge side to reduce the network transmission delay.

On the other hand, in deep learning models commonly used in medical image analysis scenarios, low-precision fixed-point computing such as INT8 can be used to reduce bandwidth bottlenecks by more efficiently utilizing caches and to improve the inference speed of models by maximizing the processor computing resources. As a result, Vistel takes full advantage of the processor features of the Intel® architecture and implements model optimization with the OpenVINO™ toolkit.

As shown in Figure 2-1-21, the OpenVINO™ toolkit converts the trained model (assuming the PyTorch framework is used) into an ONNX model using the tools provided by PyTorch, then uses the model optimizer to convert it into the optimized Intermediate Representation (IR) files unique to the OpenVINO™ toolkit that contain files in both bin and xml formats. Finally, the Calibrate Tool uses the annotation dataset to further quantify the model.

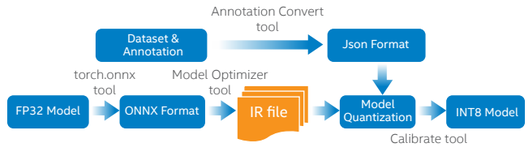


Figure 2-1-21 Model Optimization with the OpenVINO™ Toolkit

Assuming that the ResNet34 model is used, the input image resolution is **256*256**; and the task is a **NORMAL, CNV, DME**, and **DRUSEN** classification task on OCT. First, use the torch.onnx tool to convert the model into ONNX format with the following code sample:

```
SIZE = 256
dummy_input = torch.randn (8, 3, SIZE, SIZE)
model = net._dict__['resnet34'] (pretrained = True, num _ classes = 4)
model.load_state_ dict(torch.load('model/resnet34.pth', map_location = 'cpu'))
model.eval( )
torch.onnx.export(model, dummy_input, "model/resnet34.onnx", verbose = True)
```

Use the Model Optimizer tool to generate the IR file with the following command:

```
python mo_onnx.py --input_model model/resnet34.onnx --data_type FP32 --
output_dir model/FP32 --input_shape [1,3,256,256] --scale 255
```

The result is shown in Figure 2-1-22, where the IR file is in FP32 format and includes **resnet34.xml** and **resnet34.bin**.

```
Model Optimizer arguments:
Common parameters:
        - Path to the Input Model:      /home/jie.wang/code/intel_ncs/model/resnet3
4.onnx
        - Path for generated IR:        /home/jie.wang/code/intel_ncs/model/FP32
        - IR output name:       resnet34
        - Log level:    ERROR
        - Batch:        Not specified, inherited from the model
        - Input layers:         Not specified, inherited from the model
        - Output layers:        Not specified, inherited from the model
        - Input shapes:         [1,3,256,256]
        - Mean values:  Not specified
        - Scale values:         Not specified
        - Scale factor:         255.0
        - Precision of IR:      FP32
        - Enable fusing:        True
        - Enable grouped convolutions fusing:   True
        - Move mean values to preprocess section:       False
        - Reverse input channels:       False
ONNX specific parameters:
Model Optimizer version:        2019.1.1-83-g28dfbfd

[ SUCCESS ] Generated IR model.
[ SUCCESS ] XML file: /home/jie.wang/code/intel_ncs/model/FP32/resnet34.xml
[ SUCCESS ] BIN file: /home/jie.wang/code/intel_ncs/model/FP32/resnet34.bin
[ SUCCESS ] Total execution time: 6.03 seconds.
```

Figure 2-1-22 IR File Generated using the Model Optimizer Tool

The next step is to further quantify the model, which involves preparing task-related dataset and annotations, and converting the dataset into a standard format using the Annotation Convert tool provided with the OpenVINO™ toolkit. Because we use a model that performs multi-classification task, so the data is organized using the imagenet format and then converted using the tool. As shown in Figure 2-1-23, the dataset is organized in a format that includes, from left to right, image folder, image annotations, and names corresponding to the annotations.

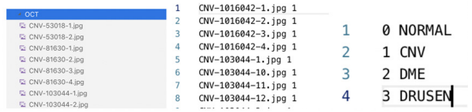


Figure 2-1-23 Organization of the Dataset

The annotation conversion command is as follows:

```
python accuracy_checker_tool/convert_annotation.py imagenet --annotation_file labels.txt --
labels_file synset_words.txt -o datasets/annotations/OCT/ -a oct.pickle -m oct.json
```

Once the conversion is complete, we obtain a json file:

```
oct.json:
{
  "label_map": {
    "0": "NORMAL",
    "1": "CNV",
    "2": "DME",
    "3": "DRUSEN"
  }
}
```

The Calibrate Tool provided with the OpenVINO™ toolkit is used to further quantify the model from FP32 to INT8 to further improve the speed of model inference. The **resnet34.yml** used in this document includes the definition and weights of the model, the types of tasks of the model, the framework used, the dataset used, etc. The file is as follows:

```
models:
  - name: resnet34
    launchers:
      - framework: dlsdk
        device: CPU
        model: resnet34.xml
        weights: resnet34.bin
        adapter: classification
        cpu_extensions: AUTO
    datasets:
      - name: OCT
```

Use **definition.yml** to define the framework and devices for launchers, as well as the address, annotation, and evaluation metrics for various datasets. Here, we use the top1 evaluation metric for accuracy. The file is as follows:

```
launchers:
  - framework: dlsdk
    device: CPU
datasets:
  - name: OCT
    data_source: datasets/OCT
    annotation: oct.pickle
    dataset_meta: oct.json
    metrics:
      - name: accuracy @ top1
        type: accuracy
```

The calibration command used by the Calibrate Tool is as follows:

```
python calibrate.py --config resnet34.yml -d  definitions.yml -
M ~/intel/openvino/deployment_tools/model_optimizer --source  ~/code/intel_ncs/ -
annotations  ~/code/intel_ncs/ --models ~/code/intel_ncs/model/FP32
```

Subsequent validation tests have shown that with the OpenVINO™ toolkit, AI applications can more fully exploit the computing resources of Intel® processors. After further conversion to the INT8 model, the inference speed can be significantly improved without affecting the accuracy, which can effectively shorten the response time of image processing and facilitate medical organizations to improve diagnosis and treatment efficiency.

Conclusion

Medical image segmentation and object detection are important branches of AI application in healthcare. A good image segmentation model can effectively facilitate medical institutions to improve the efficiency of medical image interpretation, thereby enhancing clinical diagnosis and treatment capability, improving disease cure rate, reducing patients' waiting time, and remedying various problems caused by shortage of resources in imaging departments.

Different from AI-based applications in other image processing fields, image segmentation in the medical field requires higher timeliness, after all there is only a short valuable time left for diagnosis and treatment. Therefore, if the inference efficiency of image segmentation AI application is not high enough, precious rescue may be delayed. Cases from various industries and scenarios show that products and technologies such as Intel® Xeon® Scalable Processor, 2nd Generation Intel® Xeon® Scalable Processor, Intel® DL Boost instruction set, and OpenVINO™ toolkit can effectively improve the inference efficiency of the deep learning model. With these innovative products and technologies, Intel will continue to promote and explore the innovation and implementation of AI in healthcare, thereby making technology better serve people's healthy lives.

AI + Cloud to Enable More Efficient Medical Image Analysis

Medical Image Analysis

Challenges to Medical Image Analysis

As we all know, accurately understanding and assessing illness is the prerequisite for high-quality diagnosis and treatment. In ancient times, highly skilled doctors relied on simple examinations and inquiries to understand and infer patients' illness. Today, with the help of various medical equipment and information systems, especially medical imaging equipment, doctors can better navigate through the diagnosis and treatment process and provide high-quality medical services for patients. At present, in large and medium-sized medical institutions, X-ray machines, CT scanners, MRI scanners, and so forth. have gradually become popular. Even in primary medical institutions, patients can receive various medical imaging examinations.

Although medical imaging equipment and systems can be put in place quickly, "soft power" cannot be built overnight. For example, medical imaging analysis requires doctors in the imaging department to have high degree of expertise, including knowledge in clinical medicine, medical imaging, etc., and skills in radiology, CT, MRI, ultrasound, etc. Moreover, they also need the ability to use various image analysis techniques for disease diagnosis.

Therefore, although medical imaging equipment has become quite popular in medical institutions, in some remote areas or primary medical institutions, an embarrassing situation often happens, that is, equipment has been installed but no one has the ability to review and interpret the medical images. In some provinces, for example, many medical imaging equipment have been installed in medical institutions at the county and community levels. However, after the patients have been examined, the local hospitals are still unable to make accurate judgment and analysis, and need to transmit the image files to the higher medical institutions through photographing, scanning and so on. Sometimes the quality of the image file is not guaranteed or even distorted, resulting in delay or misjudgment of the disease.

Moreover, the information systems of various medical institutions are independent of each other and the data standards are not completely unified. For example, the medical image data stored in various PACS are almost disconnected, becoming information silos. Due to the impact of these challenges, patients from remote areas can not receive effective analysis for their diseases in primary medical institutions. Even after long-distance traveling to large hospitals, they still need to undergo repeated examinations which may lead to doctor-patient conflicts.

Application of "Cloud Technology + Big Data" in Medical Image Analysis

The rapid development of cloud computing technology will gradually solve the problem of information silo. As shown in Figure 2-2-1, more and more medical institutions have begun to link medical equipment and medical service processes with cloud, and to build cloud-based capabilities and applications such as medical collaboration platform and image collaboration platform, thereby providing Platform as a Service (PaaS) or Software as a Service (SaaS) to meet different needs of various medical institutions.

Taking the full-featured medical collaboration platform as an example, medical institutions at all levels can obtain full-featured medical applications across terminals and platforms by accessing cloud services. The image collaboration platform enables medical imaging experts from large and medium-sized medical institutions to process image data from different regions anytime and anywhere, and to carry out collaborative consultation on difficult and complicated diseases, so as to achieve efficient sharing of medical resources.

Taking medical image data as an example, the interconnection and intercommunication based on cloud computing and big data technology not only enables medical institutions to avoid problems such as excessive examination and repeated treatment, but also effectively breaks data silos, provides unlimited medical services, and improves the quality of medical services. At the same time, through the accumulation and analysis of image data, the application of AI-based medical image analysis is becoming more and more mature. Now, the cloud + AI-based medical image analysis has gradually been implemented in various medical institutions and has received good feedback.

AI-based Medical Image Analysis

The massive data collected by cloud services and big data systems, provides AI models such as object detection neural networks with a large number of training samples, allowing intelligent AI-based auxiliary diagnosis systems to more effectively facilitate medical institutions improve their diagnosis and treatment capabilities.

Taking the early detection of lung cancer for an example; lung cancer is a horrible malignant tumor, while early lung cancer often appears as asymptomatic and easily neglected pulmonary nodule. Early identification

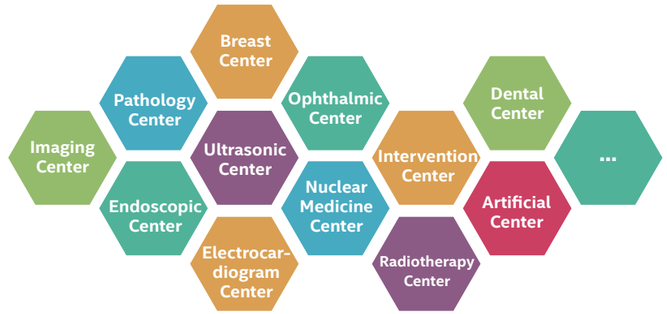


Figure 2-2-1 Link Medical Devices with Cloud Service

of pulmonary nodules (benign or malignant) can effectively reduce the mortality rate of lung cancer. Because the tiny pulmonary nodules are often difficult to be detected by human eyes in a timely and accurate manner, lung cancer, once detected, is often in the middle and late stages, resulting in the optimal treatment window no longer being available.

Now, in the AI-based medical image analysis application, as shown in Figure 2-2-2, some medical institutions are using low-dose CT to carry out intelligent auxiliary diagnosis of pulmonary nodules. Practical data shows that its quantitative monitoring sensitivity (detection rate) has reached 95%, and the screening time has been shortened from more than 10 minutes required by manpower to seconds[19]. After pulmonary nodules are identified by AI model, they are then subject to further diagnosis by doctors, so the efficiency and accuracy are greatly improved.

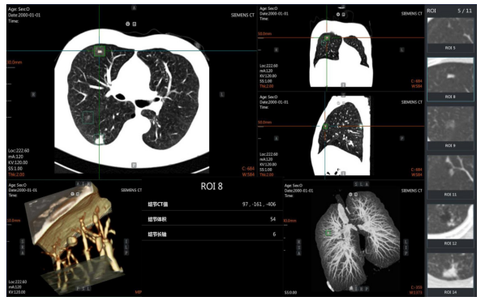


Figure 2-2-2 Intelligent Auxiliary Diagnosis of Pulmonary Nodules with Low Dose CT

At present, object detection neural network is being widely used in AI-based medical image analysis application, which can efficiently and accurately detect lesions on X-ray images, CT images and other medical images by using deep learning technologies.

Object Detection Neural Network

Typical object detection neural networks include R-CNN, Fast R-CNN, SPP-NET, R-FCN[20] and so on. In recent years, R-FCN is a widely used object detection neural network model in medical image analysis applications.

A typical R-FCN structure is shown in Figure 2-2-3. Firstly, the images to be processed are preprocessed and then sent to a pre-trained Convolutional Neural Network (CNN), such as ResNet-101 network. On the feature map obtained at the last convolutional layer of the network, three branches will be generated. The first branch will import the feature map into the Region Proposal Network (RPN) and obtain the corresponding Region of Interest (ROI). The second branch will obtain a multidimensional position-sensitive score map for classification on the feature map. The third branch will obtain a multidimensional position-sensitive score map for regression on the feature map. Finally, on the two position-sensitive score maps, Position-Sensitive ROI Pooling operations will be performed to obtain corresponding category and position information.

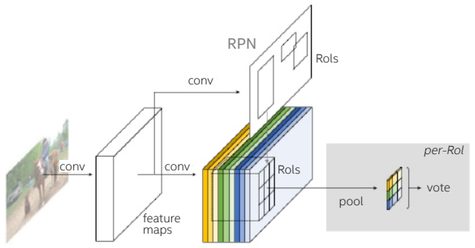


Figure 2-2-3 Typical R-FCN Structure

Compared with other object detection neural network models, such as Faster R-CNN, R-FCN has faster detection speed and higher detection accuracy[21].

Recommended Hardware and Software Configuration

When building an AI-based medical image analysis solution, please refer to the following hardware and software configurations based on Intel® platform.

Name	Specification
Processor	Intel® Xeon® Gold 6240 processor or higher
Hyper Threading	On
Tutbo Boost	On
Memory	16GB DDR4 2666MHz* 12 and above
Storage	Intel® SSD D5 P4320 Series and above
Operating System	CentOS Linux 7.6 or latest version
Linux Kernel	3.10.0 or latest version
Compiler	GCC 4.8.5 or latest version
Caffe Version	Intel® Optimization for Caffe 1.1.6 or latest version

[19] Data is quoted from Accurad's internal test data: https://www.intel.cn/content/www/us/en/analytics/artificial-intelligence/yinggu-case-study-medical.html
[20] R-FCN related technical description is quoted from Jifeng Dai, Yi Li, Kaiming He, Jian Sun, R-FCN: Object Detection via Region-based Fully Convolutional Networks, https://arxiv.org/pdf/1605.06409v2.pdf
[21] For R-FCN performance data, please refer to Jifeng Dai, Yi Li, Kaiming He, Jian Sun, R-FCN: Object Detection via Region-based Fully Convolutional Networks, https://arxiv.org/pdf/1605.06409v2.pdf

Optimizing AI Model Efficiency

Optimizations Based on Intel® Processor Platforms

Intel® processor platforms, including Intel® Xeon® Scalable Processor, 2nd Generation Intel® Xeon® Scalable Processor, can not only bring powerful general-purpose computing power to intelligent AI + Cloud-based medical image analysis systems but also provide much-needed parallel computing power. The inference of deep learning model often requires higher parallel computing capability. By introducing Intel® AVX-512, Intel® Xeon® Scalable Processor provides more efficient Single Instruction Multiple Data (SIMD) execution and enables the system to obtain more powerful parallel computing acceleration capability.

In addition, the newly added Intel® Math Kernel Library (Intel® MKL) and Intel® MKL-DNN can further improve the working efficiency of AI models by enhancing the performance of AI models in the following three aspects:

- Use Cache Blocking to optimize data cache and improve data hit rate;
- Perform parallelization and vectorization optimization on common operators in neural networks;
- Use Winograd algorithm-level optimization.

Intel® DL Boost included with the new 2nd Generation Intel® Xeon® Scalable Processor enables deep learning inference to use INT8 for better performance.

For example, when using R-FCN model to process a single thoracic Dicom data on the Intel® Xeon® Scalable Processor platform, data from an application show that the optimized Intel® Xeon® Gold 6148 processor can improve performance by nearly 5 times[22].

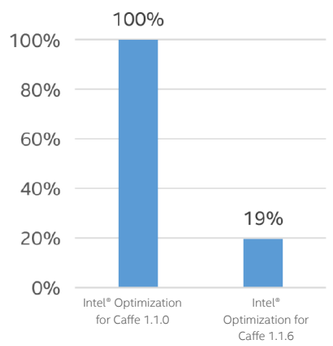


Figure 2-2-4 Comparison When Using R-FCN Model to Process a Single Thoracic Dicom Data

Intel® Optimization for Caffe

Compared with the Berkeley Vision and Learning Center (BVLC) distribution of Caffe[23], Intel® Optimization for Caffe[24] is specially optimized for Intel® architecture and provides support for Intel® MKL, Intel® MKL-DNN and Intel® AVX-512, which has better performance and higher inference efficiency in all deep learning models.

In order to make full use of the computing resources provided by Intel® processors, it is a common practice to set some environment variables before performing inference, such as:

```
1. export OMP_NUM_THREADS=36
```

Here, **OMP_NUM_THREADS** is to specify the number of threads to use.

After analyzing the performance of BVLC Caffe, users can find that Intel® Optimization for Caffe has been optimized in the following aspects.

■ Code Vectorization Optimization

The optimization includes:

- The Basic Linear Algebra Subroutine (BLAS) library is switched from the Automatically Tuned Linear Algebra Software (ATLAS) to Intel® MKL-DNN, thus making the General Matrix Multiply (GEMM) more suitable for vectorized and multi-threaded workloads and increasing the cache size.
- Use the Xbyak Just-in-Time (JIT) assembler to perform the compilation process. As an x86/x64 JIT assembler, Xbyak provides better support for instruction sets under Intel architecture, such as MMX™ technology, Intel® Streaming SIMD Extensions (Intel® SSE), and Intel® AVX series technology. At the same time, it can also help Intel® Optimization for Caffe to improve vectorization rate during code implementation.
- Perform code vectorization for GNU Compiler Collection (GCC) and Open Multi-Processing (OpenMP). The improvement of vectorization rate will facilitate **SIMD** instructions to process more data at a time and improve data parallel utilization rate. In addition, code vectorization can also effectively improve the performance of pooling layer in deep learning model.

[22] The performance test results are based on the tests performed on April 10, 2019, and the test configuration is as follows: 2S Intel® Xeon® Gold 6148 Processor, 20 cores/40 threads, HT/Turbo enabled, 192GB of memory (12 slots/16GB/2666MHz), CentOS 7.6, BIOS:SE5C620.86B.02.01.0008.031920191559 (uncode:0x200005e), Kurnel version: 3.10.0-957.21.3.el7.x86_64, compiler GCC 4.8.5. The test group uses Intel® MKL-DNN version 0.12, while the control group uses Intel® MKL-DNN version 0.18; Framework: the test group uses Intel® Optimization for Caffe 1.1.0, while the control group uses Intel® Optimization for Caffe 1.1.6. Minibatch=1.
[23] For the source code of this version, please refer to https://github.com/BVLC/caffe.
[24] For the source code of this version, please refer to https://github.com/intel/caffe.

■ General Code Optimization

The optimization includes:

- Reduce programming complexity;
- Reduce the number of calculations;
- Expand the loop.

For example, some scalar optimization techniques are used in the code optimization process. The code is as follows:

```
1.  for (int h_col = 0; h_col < height_col; ++h_col) {
2.    for (int w_col = 0; w_col < width_col; ++w_col) {
3.      int h_im = h_col * stride_h - pad_h + h_offset;
4.      int w_im = w_col * stride_w - pad_w + w_offset;}}
```

In the third line of its code fragment, the **h_im** calculation can be moved out of the innermost layer as follows:

```
1.  for (int h_col = 0; h_col < height_col; ++h_col) {
2.    int h_im = h_col * stride_h - pad_h + h_offset;
3.    for (int w_col = 0; w_col < width_col; ++w_col) {
4.      int w_im = w_col * stride_w - pad_w + w_offset;}}
```

■ Other Optimizations based on Intel® Processors

The optimization includes:

- Improve execution efficiency of **im2col_cpu/col2im_cpu**; **im2col_cpu** is a commonly used function in deep learning computing, which can use optimized BLAS library to perform direct convolution in GEMM mode. The following optimizations can be implemented for **im2col_cpu**: in **BVLC Caffe** code,

```
1.  for (int c_col = 0; c_col < channels_col; ++c_col)
2.    for (int h_col = 0; h_col < height_col; ++h_col)
3.      for (int w_col = 0; w_col < width_col; ++w_col)
4.        data_col[(c_col*height_col+h_col)*width_col+w_col] = // ...
```

four arithmetic operations (two additions and two multiplications) can be replaced by a single index increment operation to improve the operation efficiency.

- Reduce the complexity of normalized batch processing;
- Optimizations for specific processors/systems;
- Each computing thread locks one core to avoid thread movement, which can be achieved by setting the following environment variables.

```
1.  export KMP_AFFINITY=granularity=fine,compact,1,0
```

By setting adjacent threads, GEMM operation performance can be improved. Because all threads can share the same Last Level Cache (LLC), so the previously pre-fetched cache lines can be reused for data, improving efficiency.

■ Code Parallelization with OpenMP

The OpenMP multithreading parallel processing method can effectively improve the inference efficiency of the neural network. For example, in the pooling layer, a single pooling layer is suitable for processing a single feature map. However, if the pooling layer and OpenMP multithreading are executed in parallel, because the images are independent of each other, multiple threads can process multiple images simultaneously in parallel to improve the efficiency. The code is as follows:

```
1.  #ifdef _OPENMP
2.    #pragma omp parallel for collapse(2)
3.  #endif
4.    for (int image = 0; image < num_batches; ++image)
5.      for (int channel = 0; channel < num_channels; ++channel)
6.        generator_func(bottom_data, top_data, top_count, image, image+1,
7.                       mask, channel, channel+1, this, use_top_mask);
8.  }
```

It can be seen that with the **collapse(2)** clause, OpenMP **#pragma omp parallel** can be extended to two for-loop nested statements, and then the two loops of batch iteration image and image channel are merged into one loop and parallelized.

With the aid of a series of optimization methods and techniques, the performance of Intel® Optimization for Caffe has been greatly improved compared with BVLC Caffe. A test shows that, compared with the native Caffe, Intel® Optimization for Caffe can reduce the workload execution time by 10% and improve the overall execution performance by more than 10 times[25].

*For more technical details of Intel® Optimization for Caffe, please refer to the relevant content in the Technologies section of this manual.

[25] For relevant test data and more optimizations of Intel® Optimization for Caffe, please refer to **Caffe* Optimized for Intel® Architecture: Applying Modern Code Techniques:** https://software.intel.com/en-us/articles/caffe-optimized-for-intel-architecture-applying-modern-code-techniques.

Xi'an Accurad Utilizes AI and Cloud Services to Aid in Medical Diagnosis and Treatment

Background

The unbalanced allocation of medical resources has led to varying post-processing and post-analysis capabilities of medical images in various medical institutions. In addition, the lack of data connectivity also makes it difficult to effectively improve the utilization efficiency of medical resources through resource sharing. Xi'an Accurad Network Technology Co., Ltd. (hereinafter referred to as "Xi'an Accurad"), which has been focusing on medical imaging core technologies for nearly 20 years, is committed to combining its professional medical imaging core technologies and products with the latest cloud computing, big data and AI technologies to build an efficient and intelligent medical auxiliary diagnosis capability and to facilitate medical institutions to improve diagnosis and treatment efficiency and quality.

According to Xi'an Accurad, to solve the problem of unbalanced development of medical image analysis and processing capabilities, it's necessary to collect medical image data by using cloud computing and other methods, build AI-driven data analysis capabilities based on such data, and then gradually eliminate the differences in medical image analysis capabilities of different medical institutions by means of resource sharing and AI.

For this reason, Xi'an Accurad, by implementing the Yizhen cloud, use its innovative AMOL technology which is an Internet of Things implementation consisting of various medical technologies and equipment, to link massive medical image data from different equipment. In addition, Xi'an Accurad also introduces deep learning into medical image processing, and build a new Cloud IDT service based on the object detection neural network model, which has achieved remarkable results in improving detection rate, reducing decision-making time and improving work efficiency.

In order to facilitate Xi'an Accurad to better implement this system, Intel has provided it with the latest generation of platform products and technologies such as Intel® Xeon® Scalable Processor, facilitating it complete the migration of Cloud IDT service to Intel® platform as well as the deployment and optimization of deep learning frameworks such as Caffe and TensorFlow. Thanks to the cooperation and efforts of both parties, the performance of the brand-new intelligent medical auxiliary diagnosis system in several indicators such as screening time and accuracy rate has won unanimous praise from customers.

Solution and Results

In the new solution, Xi'an Accurad develops a series of medical image analysis and processing applications based on the object detection neural network model, and uses Intel® processors to carry out efficient model inference. In addition, Xi'an Accurad also integrates its intelligent Cloud IDT applications with @iMAGES, a medical image processing and analysis cloud, to provide powerful intelligent online processing capability for images and big data. As shown in Figure 2-2-5, by combining the strong computing power provided by Intel® processors with the cloud-based powerful Positron Emission Tomography CT (PET-CT) fusion capability provided by @iMAGES core engine, this solution can not only provide the "heat map" based on morphology and function, but also perform semi-quantitative Standard Uptake Value (SUV) analysis on the images which can be subject to further identification and quantitative analysis of diseases such as tumors by the R-FCN object detection neural network in the intelligent Cloud IDT system.

In addition to excellent hardware performance, Intel has further improved the execution efficiency of Xi'an Accurad's Cloud IDT intelligent system by optimizing AI frameworks such as Caffe and TensorFlow. Thanks to the optimization of R-FCN model, the model clipping and fusion has improved the performance by nearly 30%. After further optimizing the OpenMP multi-threading implementation, the performance has improved by 40-50%[26].

Further, the general-purpose computing power and parallel computing power provided by Intel® Xeon® Scalable processors can also enable the intelligent system to merge tasks previously distributed on different platforms such as data statistics and model inference tasks, thus allowing users not only to deploy more virtual machines in their private cloud, but also to reduce the Total Cost of Ownership (TCO).

Now, Xi'an Accurad has developed a series of intelligent auxiliary diagnostic capabilities such as pulmonary nodule diagnosis, rib fracture diagnosis and tuberculosis diagnosis based on the AI + Cloud model, some of which are shown in the following table:

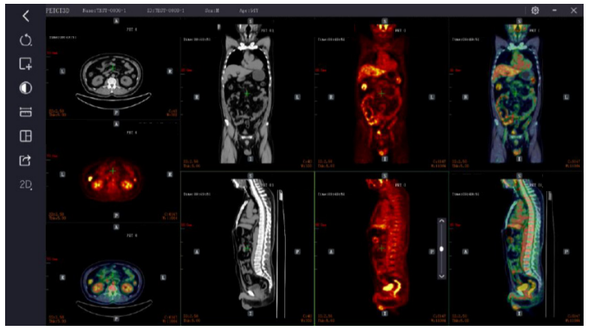


Figure 2-2-5 Cloud-based PET-CT Fusion

[26] Accurad's internal test data: https://www.intel.cn/content/www/cn/zh/analytics/artificial-intelligence/yinggu-case-study-medical.html, the test configuration is as follows: Processor: 2S Intel® Xeon® Gold 6148 Processor, 2.40GHz; Cores/Threads: 20/40; HT: ON; Turbo: ON; Memory: 192GB DDR4 2666; Hard Disk: Intel® SSD SC2KB48; Network Adapter: Intel® Ethernet Converged Network Adapter XC710; BIOS: SE5C620.86B.02.01.0008.031920191559; Operating System: CentOS Linux 7.6; Linux Kernel: 3.10.0-957.21.3.el7.x86_64; gcc Version: 4.8.5; Caffe Version: Intel® Optimization for Caffe 1.1.6; Workload: R-FCN.

Intelligent Auxiliary Diagnosis System Built by Xi'an Accurad on AI + Cloud[27]	
Diagnosis of Pulmonary Nodule	On the basis of chest CT data labeled by a large number of senior doctors, a specific deep neural network and image algorithm, which is designed by using deep learning technology and 3D image processing technology, can locate pulmonary nodules over 3mm from chest CT data and calculate the nodule size and the nodule malignancy risk, with a detection accuracy up to 95%.
Diagnosis of Rib Fracture	Focusing on the detection of rib fracture, the fully intelligent automatic detection system built with chest x-ray image data, uses deep learning technology and image processing technology to automatically identify and locate fractures and automatically label them on images, with a detection accuracy above 90%, thereby helping doctors to quickly find and diagnose such fractures.
Diagnosis of Tuberculosis	Built with advanced image processing and AI machine learning algorithm, the automatic tuberculosis detection system for chest X-ray image can scan high-resolution and digital chest X-ray images, automatically detect and score the suspicious lesions, and easily and quickly display the detection results, with sensitivity as high as 86%, providing doctors with useful diagnostic information.
AI-based Pneumonia Detection	By detecting the location of suspected pneumonia lesions, the AI-based pneumonia detection for chest X-ray images can be used for daily screening of common pulmonary inflammatory diseases and can help doctors improve their diagnostic efficiency. Its sensitivity of pneumonia detection is up to 82%.
Healthy Chest X-ray Image Screening	The healthy chest X-ray image screening is designed to screen out healthy chest X-ray images from all chest X-ray images, which can reduce the screening workload and help doctors focus on the abnormal data. This service is mainly used for screening scenarios (such as physical examination), and its sensitivity and specificity can reach 99% and 22% respectively when screening healthy chest X-ray images.
Intelligent Screening of Pneumoconiosis	Built with advanced image processing and AI machine learning algorithm, the automatic pneumoconiosis detection system for chest X-ray image which is designed by using specific deep neutral network and imaging algorithm, can scan high-resolution and digital chest X-ray images, automatically detect the suspicious lesions, and easily and quickly display the detection results, with sensitivity as high as 92%.
Intelligent Early Screening for Breast Cancer	A specifically designed deep neural network, which is built by deep learning molybdenum target X-ray images of breast labeled by a large number of experts, can automatically identify abnormal masses, calcified lesions, etc. in molybdenum target X-ray images, with over 95% of sensitivity for detecting calcified spots and masses, and 92% of sensitivity for identifying calcification.
Screening of Diabetic Retinopathy	Fundus photography is a fully intelligent automatic detection system that uses fundus data to detect and predict different levels of diabetic retinopathy. By building data verification and medical logic modules, it has the ability to recognize various fundus lesions and diseases in fundus images and assist doctors in effectively screening out early patients, thereby reducing misdiagnosis and missed diagnosis.
Intelligent Recognition of Small Bowel Obstruction	By incorporating specific intelligent AI algorithm and imaging algorithm, the fully intelligent automatic detection of plain X-ray images of bowel obstruction is designed to detect liquid levels within the enteric cavity, recognize various obstructive lesions and diseases in abdominal erect images, provide various screening methods, and assist doctors in effectively identify patients with acute abdomen, thereby reducing misdiagnosis and missed diagnosis.
Intelligent CTA Coronary Artery Diagnosis	The fully automatic CTA coronary artery diagnosis is designed to intelligently process and analyze CTA thin-layer images through automatic coronary artery segmentation, segmentation detection, plaque classification and detection, stenosis analysis and calcification score. Combined with VR and visualized sections, it can automatically output structured report, thereby improving the efficiency of existing CTA coronary artery diagnosis by more than 6 times.

[27] Accurad AI application introduction and related data are quoted from official website of Accurad Yizhen Cloud and AI: http://ai.yizhen.cn/#Page03

Huiyi Huiying Uses a True AI Solution to Prevent and Control the COVID-19 Pandemic

Background

Imaging is a vital line of defense against epidemics and an indispensable tool to measure diagnosis, treatment and healing. As early as the beginning of 2020, the National Health Commission released the "Diagnosis and Treatment Protocol for Novel Coronavirus Pneumonia (Trial Version 6)", which clearly states that the significant imaging change shall be added as one of the criteria for determining severe cases. In the clinical practice of diagnosis and treatment of Novel Coronavirus Pneumonia, imaging plays an irreplaceable role from detecting early pulmonary abnormalities in suspected cases, confirming the cases and determining the degree of disease, and diagnosing and excluding suspected cases, to developing and adjusting treatment plans, tracking changes in the disease, and evaluating treatment efficacy and outcomes.

As a global AI player dedicated to computer vision and deep learning applications in medical imaging field, Huiyi Huiying considers the epidemic prevention and control as an overriding priority, actively engages in the fight against the epidemic by using various technologies with its consistent and rigorous attitude, and builds an intelligent AI-powered COVID-19 imaging solution to improve the diagnosis and treatment of Novel Coronavirus Pneumonia and the epidemic prevention and control.

Solution and Results

In view of the rapid spread of the epidemic, a task force of engineers and technicians worked closely in product development, upgrading as well as application, with an aim to not only race against time and compete with the epidemic, but also ensure that the products are accurate, professional and usable. Based on the overall architecture of Huiyi Huiying's intelligent imaging solution, and in response to the prevention and control needs of the COVID-19 epidemic, Huiyi Huiying quickly developed an intelligent medical imaging and diagnosis solution, which uses AI algorithms to analyze Computed Tomography (CT) chest scan results, and to detect pneumonia lesions and give the probability of suspected COVID-19 case, thereby effectively supplementing the standard laboratory tests. The CT scan image-based and AI-assisted COVID-19 diagnosis and screening system, incorporates a series of upgrades to the original solution from algorithm to platform.

In terms of algorithm, compared with some similar products that use CT value, the solution adopts a different decision-making criterion. To this end, Huiyi Huiying team quickly collected a large amount of Novel Coronavirus Pneumonia data that is precisely annotated by a team of professional doctors, and combined them with deep learning algorithm to achieve accurate segmentation and measurement of pneumonia lesion area, thereby providing effective parameters for patients' prognosis, and predicting the type of pneumonia to facilitate clinical diagnosis and treatment.

The combination of precisely annotated data and deep learning algorithm helps determine the precise contour and volume of each individual lesion, and plays an important role in AI-assisted doctor diagnosis, thanks to its ability to precisely locate and differentiate lesions. At the same time, this AI-based solution can automatically adapt to images from different hospitals and different devices and with different layer thicknesses, and achieve self-iteration and model tuning, providing a very high detection rate and detection accuracy for pneumonia lesions. Meanwhile, by relying on Intel's powerful processors, lightweight network models and other technologies, the solution can further enhance algorithm efficiency, computing more than 500 CT images in just 2-3 seconds[28].

In terms of functions and features, the Huiyi Huiying solution can provide quantitative data on lesion location, size, area change, increase/decrease, addition/disappearance, and degree of criticality, which can effectively support doctors to efficiently and accurately evaluate patients' conditions

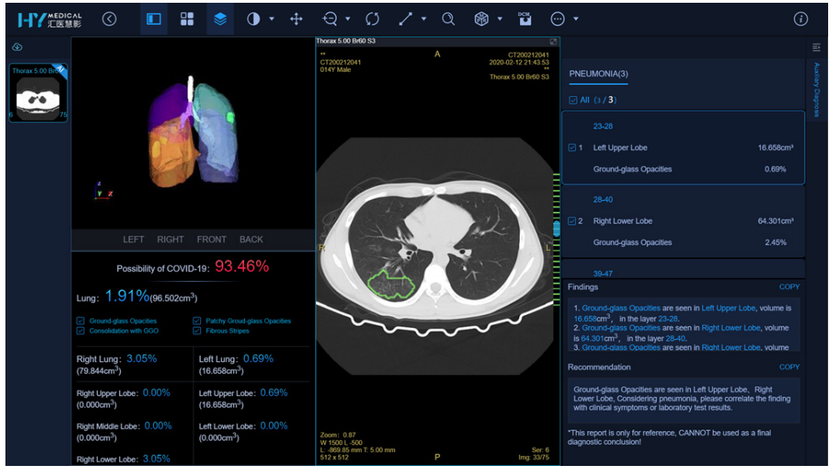


Figure 2-2-6 Early Detection of COVID-19 Infection Using CT Chest Scan

[28] Data source: https://www.leiphone.com/news/202002/2Q8aKrElPbqboY2R.html?viewType=weixin

and treatment efficacy. At the same time, combined with the latest COVID-19 prevention and control guidelines, the solution is also able to provide interactive structured reports for COVID-19 CT scan, thereby realizing structured, intelligent and standardized image reports, improving the hospital informatization and the quality of image report, facilitating data acquisition for individual diseases, as well as providing a data warehouse for further mining of these data in the future.

With the aid of advanced algorithms and field data, the Huiyi Huiying solution enables rapid screening of pneumonia signs and symptoms, labeling of suspected cases, automatic location and precise segmentation of lesions, quantitative analysis of lesions, fully automatic comparison of previous and current images, follow-up visit management, structured reporting, and so on, thereby assisting in the diagnosis and treatment of Novel Coronavirus Pneumonia throughout the whole process. In addition, through collaboration with Intel, the solution uses hardware and software based on the Intel® architecture to dramatically improve the system's inference and analysis performance, helping doctors to more quickly determine suspected cases and assess disease progression, and effectively easing the pressure on hospitals.

As a member of Intel AI Builders Program, Huiyi Huiying worked closely with Intel to implement several optimizations during the development of this solution. The solution uses the second-generation Intel® Xeon® Gold 6252N processors as the computational engine for training and inference, which have more processor cores and threads and a fully optimized and upgraded micro-architecture that can bring more computing power and provide faster inference speed.

In addition, the OpenVINO™ toolkit is used to further accelerate the operational efficiency of AI workloads. Combined with the model optimizer, instruction set optimizations and other features provided with the OpenVINO™ toolkit, especially the deep learning framework runtime acceleration capability provided by Intel® MKL-DNN, the inference performance of the solution is greatly improved. As shown in Figure 2-2-7, compared with the pipeline running on PyTorch 1.5.1, the pipeline optimized with OpenVINO™ toolkit resulted in a 2.89x performance improvement[29].

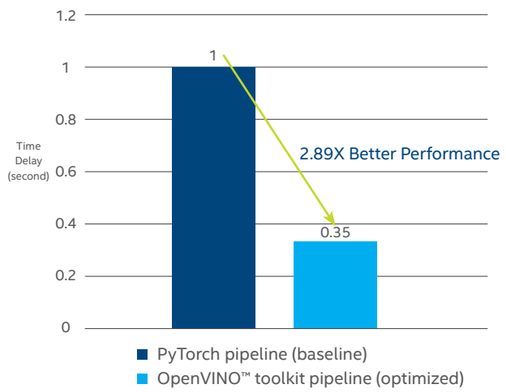


Figure 2-2-7 Benchmarking Results on Intel® Xeon® Gold 6252N Processor

Huiyi Huiying's AI-assisted COVID-19 diagnosis system demonstrates the computing power and AI acceleration capability of Intel® Xeon® platform, and is optimized with the OpenVINO™ toolkit for even greater AI inference performance. Thanks to its excellent performance, the system was highly praised after its implementation in Beijing You'an Hospital affiliated with Capital Medical University, helping doctors to improve the speed and accuracy of COVID-19 diagnosis, clearly displaying the lesion location and lesion data, and enabling automatic and quantitative comparison for the follow-up of COVID-19 patients' lesions, which is the key of epidemic prevention and control. The system also gave a clear picture of disease progression to meet the needs of actual clinical work, truly making AI an effective aid to the epidemic prevention and control on the frontlines.

Today, the system has been rapidly implemented in more than 50 hospitals fighting COVID-19 on the frontlines in countries including China, the United Kingdom, Italy, Belgium, Mexico, Chile, Ecuador, Panama, etc., becoming an exemplary solution for the prevention and control of COVID-19 with artificial intelligence technology.

[29] The configuration of the performance test is as follows: 2S Intel® Xeon® Gold 6252N Processor @ 2.30GHz, 24 Cores, Turbo On, HT On, BIOS 4.1.13, Memory 192GB, OS: Ubuntu 18.04.4 LTS. OpenVINO R2020.3.194.

Conclusion

The bright future of data-driven medical informatization is the common aspiration of Intel, Xi'an Accurad and other partners. By utilizing cloud computing, Internet of Things, big data, and AI, digitized and intelligent healthcare applications have been widely implemented and developed, breakthroughs have been made in real-time computation and display of medical image data, AI research of medical visual data, and tangible results have been obtained from the actual deployment and implementation in various medical institutions.

In order to continuously tap the potential of current mainstream AI frameworks on Intel® platforms, Intel has carried out multi-faceted optimizations for these frameworks. Intel® Optimization for Caffe has greatly improved the overall performance of the model compared with BVLC Caffe through code vectorization, OpenMP parallelization, etc. After combining with Xi'an Accurad's applications such as intelligent Cloud IDT application cloud-based medical image processing and analysis and @iMAGES core engine, the intelligent "AI+Cloud" auxiliary diagnostic capabilities have been built in numerous key scenarios such as pulmonary nodule diagnosis. With the optimizations provided by OpenVINO™ toolkit, the AI inference performance of Huiyi Huiying's AI-assisted COVID-19 diagnosis system has also been greatly improved.

With the launching of the new generation of Intel technologies and products such as the 2nd Generation Intel® Xeon® Scalable Processor and Intel® Optane™ Persistent Memory, it is expected that medical imaging analysis solutions built on Intel® platforms will provide more powerful performance and superior AI capabilities. In the future, Intel also plans to continue its in-depth cooperation with more partners to incorporate more advanced products and technologies into the process of medical informatization, thereby promoting the development of precise and intelligent healthcare, effectively improving the level of medical services with informatization, digitization and intelligence, and providing patients with more comfortable and personal medical and health services.

AI Accelerates Pathological Image Analysis

Pathological Section Analysis in Medical Practice

Challenges of Traditional Pathological Section Analysis

Pathology is the process that prepares micron-sized sections of pathological tissues or organs in several steps, adheres them to glass slides, dye them, and then hand them over to pathology department where pathologists will examine pathological sections under microscope to observe pathological changes and make pathological diagnosis and prognosis evaluation. Pathological section examination is a very complicated and challenging job. Becoming an expert in pathology requires many years' experience in observing tens of thousands of samples as well as rich professional knowledge. However, according to statistics, there are currently less than 10,000 pathologists in China[30].

In addition, manual examination is inevitably subjective, and the diagnosis made by different pathologists on the pathological sections of the same patient often differs, which may lead to misdiagnosis and missed diagnosis. Furthermore, in the actual pathological section examination, when the patient's pathological section is digitized at a magnification of 40 times, the number of pixels of a single pathological section may exceed one million. Pathologists need to continuously observe multiple megapixel images and pay attention to the abnormalities of microscopic areas in the images, which is not only time-consuming and laborious, but also prone to errors and omissions. Moreover, a longer observation time will also result in a long waiting time for patients, which may delay the treatment of their disease.

AI-based Pathological Section Analysis

With the rapid development of AI-based image processing and analysis technology, various medical institutions have spared no effort to develop deep learning or machine learning-based pathological section analysis and achieved good results. For example, a deep learning model trained by ResNet50 network can be used to identify tumor from pathological tissues. Although noise and other problems still exist in the obtained tumor prediction heat map, pathological section images can already be examined with different magnifications like pathologists. Tests show that it is possible for medical institutions to train a deep network model which not only boosts professional detection technology, but also provides ultra-fast detection speed and unlimited working time.

A latest research from New York University shows that, the Inception v3 deep learning model, which is trained by a large number of digital pathological section images, has achieved 99% of accuracy in differentiating the tumor from normal tissues and 97% of accuracy in distinguishing adenocarcinoma and squamous cell carcinoma[31].

Now, CNN-based classification algorithms and object detection algorithms have made great progress. In addition, many models derived from CNN, one of the representative models of deep learning, such as LeNet, ZFNet, VGGNet and ResNet, have been widely used in image classification, portrait recognition, target location and image analysis.

■ Classified Convolutional Neural Network

The detection results of medical images can often be expressed as a classification problem, such as classifying negatives as normal and positives as abnormal. Now, the expected detection results are a series of discrete numbers, such as 0 or 1, which constitutes a typical classification problem. This makes us believe that, by using classification algorithms such as binary classification, CNN can effectively facilitate medical institutions to qualitatively screen out problematic areas or tissues first, and then carry out quantitative analysis and interpretation.

A typical binary classification algorithm, such as logistics regression, is a generalized linear regression analysis model. For example, when predicting whether a patient suffers from cancer based on pathological section images, it is assumed that with the increase of the patient's age, cancer can be determined when more than x cells are found. Now, the determination of cancer can be mathematically expressed as a linear function with a threshold value of **x**, i.e., **y = age(n)*a + initial value(b)**. When **y >= x**, cancer will be determined.

In actual scenarios, this function will be much more complicated. For example, in addition to age, the size and state of abnormal cells may also become the criteria for the determination of cancer. At this time, the linear function will become a multivariate linear function, like

y=n*a+m*c+o*d……+b

As mentioned earlier, the classification problem requires a series of discrete outputs, so it's necessary to add an activation function to the linear function to discretize the output. In the neural network, the role of activation function is to introduce non-linearity into the neural network so that the neural network can better solve more complex problems. Common activation functions include Sigmoid function, tanh function, ReLU function, and so on. In addition, logistics regression will use gradient descent iterative method to obtain the minimum loss function.

Generally, as shown in Figure 2-3-1, CNN binary image classification algorithm has the following key modules, including image reading and preprocessing, image training, iterative optimization and image prediction. Among them, the CNN-based model training consists of convolutional layer, pooling layer and fully connected layer and can employ cross entropy loss function, MBGD gradient descent algorithm or BGD gradient descent algorithm.

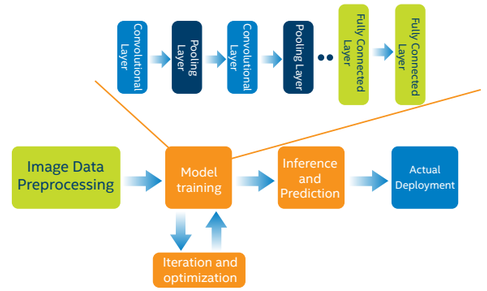


Figure 2-3-1 Modules of CNN Binary Image Classification Algorithm

[30] This data is quoted from media report: https://www.cn-healthcare.com/article/20141118/content-463705.html
[31] Source: Coudray N, Moreira A L,Sakellaropoulos T,et al.Classification and Mutation Prediction from Non-Small Cell Lung Cancer Histopathology Images using Deep Learning[J].bioRxiv, 2017.

In actual application, the Residual Net (ResNet) is also one of the common Classified convolutional neural networks, which has very excellent performance in 2D image classification, detection and localization. Compared with other CNN, ResNet adds a shortcut connection in the network, allowing input information to be directly transmitted to the following layers, as shown in Figure 2-3-2:

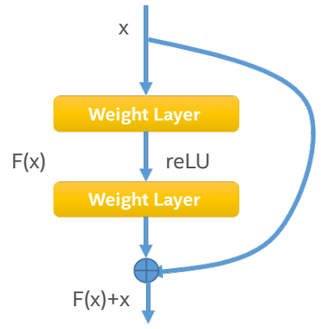


Figure 2-3-2 Residual Structure of ResNet

This structure (residual structure) partially solves the problems of information loss and even gradient disappearance that may exist in the transmission of information in the classical CNN network structure. These problems are one of the reasons why the number of layers in the deep model cannot be very large. After employing ResNet, the number of layers for training model can be greatly increased, thus improving the classification accuracy.

■ Object Detection Neural Network

Object detection neural network is designed to accurately find the location of an object in a given image and label the category of the object. Common object detection neural networks include R-CNN, Fast R-CNN, SPP-NET, R-FCN, and so forth.

R-CNN is a classical deep learning object detection algorithm. Its basic workflow is as follows: First, R-CNN will generate thousands of candidate regions with the same size on the original image by using the selective search method and input them into CNN network. The eigenvectors obtained from the network model will pass through multi-category Support Vector Machines (SVM) classifiers, and each object will train an SVM classifier and infer the probability that it belongs to the object according to the eigenvectors. At the same time, R-CNN also sets up a bounding box regression model to improve the positioning accuracy by optimizing the accurate position of the bounding box.

In order to solve the problems of slow training, inference and test speed and larger required training space in actual R-CNN applications, Fast R-CNN adopts the following techniques to address these problems and achieves better performance than R-CNN. These techniques are as follows:

- Normalize the whole image before sending it to CNN network.
- The convolutional layer does not perform feature extraction from candidate regions, but adds coordinate information of candidate regions to the last pooling layer to perform feature extraction calculation.
- The object and candidate box regressions are both performed in CNN network.

However, the subsequent Faster R-CNN integrates feature extraction, proposal extraction, bounding box regression (rect refine) and classification into one network, thereby greatly improving the overall performance, especially in terms of detection speed.

Recommended Hardware and Software Configuration

When building an AI-based pathological section analysis solution, please refer to the following hardware and software configurations based on Intel® platform.

Name	Specification
Processor	Intel® Xeon® Gold 6240 processor or higher
Hyper Threading	On
Tutbo Boost	On
Memory	16GB DDR4 2666MHz* 12 and above
Storage	Intel® SSD D5 P4320 Series and above
Operating System	CentOS Linux 7.6 or latest version
Linux Kernel	3.10.0 or latest version
Compiler	GCC 4.8.5 or latest version
Caffe Version	Intel® Optimization for Caffe 1.1.6 or latest version

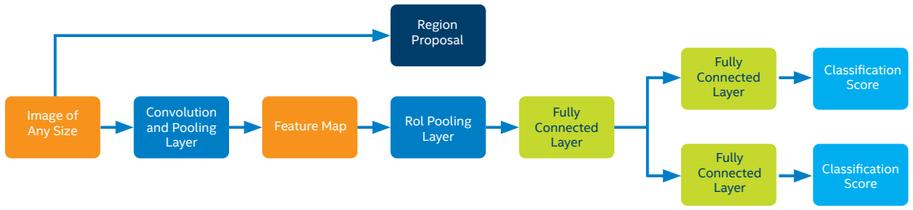


Figure 2-3-3 Structure of Fast R-CNN Network

Optimizations for Deep Learning-based Pathological Section Analysis

Optimizations Based on Intel® Processors

When building and optimizing deep learning-based pathological section analysis solution on Intel® processors, users can benefit from the following advantages:

- The file size of each pathological section image is often tens or hundreds of MB. Traditionally, due to the limitation of storage space, the Batch Size set in training is small, which will lead to an increase in training time. However, with the Intel® processor platform, the server is equipped with large memory (several TB or even dozens of TB), which allows users to set Batch Size to over 100, thereby accelerating the training speed.
- The introduction of Intel® Optane™ Persistent Memory built with 3D XPoint™ storage media further enhances the advantages of Xeon Scalable platforms. Compared with expensive Dynamic Random-Access Memory (DRAM), Intel® Optane™ Persistent Memory has large capacity and non-volatile advantages and it requires lower cost when expanding capacity, which can effectively improve the memory density and computing efficiency of servers performing model training and inference, and greatly reduce TCO.
- The innovative microarchitecture of Intel® Xeon® Scalable Processor, including more cores, more concurrent threads and more cache, together with a large number of hardware enhancements integrated with it, especially Intel® AVX-512, can provide more computing power for AI applications.

* For more technical details about Intel® Optane™ Persistent Memory, please refer to relevant content in the Technologies section of this guide.

Intel® Optimization for Caffe

Caffe is a commonly used deep learning framework, which is widely used in AI training and inference in video and image processing fields. In order to further improve and optimize the working efficiency of Caffe-based deep learning model, Intel has made a large number of optimizations to Caffe based on the features of Intel® architecture.

These optimizations include:

■ Optimization for Typical ResNet Network

Intel® Optimization for Caffe utilizes features of a series of ResNet models to reduce the overhead of computing and memory access. Figure 2-3-4 shows a typical residual structure of ResNet. As can be seen from the left half of the figure, only half of the activation operation is consumed by the two 1*1 Stride-2 convolutional layers at the bottom. The optimized structure changes the setting of binding layer, as shown in the right half of the figure, which adds a 1*1 pooling layer to the shortcut connection, reducing the computation by half.

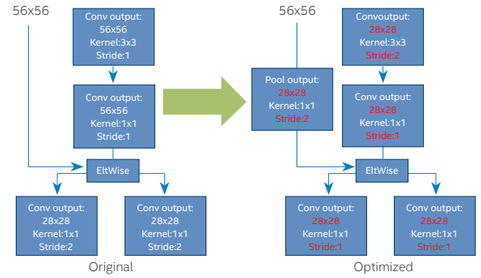


Figure 2-3-4 Optimization for ResNet with Intel® Optimization for Caffe

■ Layer Fusion Technology

In addition to optimizations for instruction set vectorization and thread-level parallelism, Intel® Optimization for Caffe also introduces more effective layer fusion optimizations into Caffe framework, such as BN+Scale, Conv+Sum, Conv+Relu, BN inplace and sparse fusion, which greatly improve the performance of neural networks, such as ResNet50. As shown in Figure 2-3-5, this is a fusion of Conv layer and Eltwise layer with residual structure. The Conv layer res2a_branch2c and Eltwise layer res2a_relu in the left half of the figure are fused into a new Conv layer res2a_branch2c (shown in the right half of the figure), which effectively improves the performance of ResNet network model.

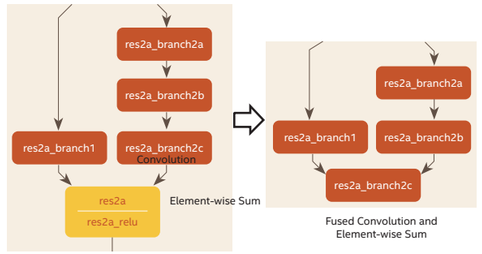


Figure 2-3-5 Fusion of Conv Layer and Eltwise Layer

In addition, Intel® Optimization for Caffe also provides better support for INT8, and provides a calibration tool to facilitate users seamlessly convert neural networks to INT8 to greatly improve performance.

A test shows that, compared with BVLC Caffe, when performing AI inference on Intel® Xeon® Scalable Processor by using ResNet50 convolutional neural network in the same evaluation environment, Intel® Optimization for Caffe, by adding layer fusion technology, increases the inference performance per unit time by 51 times while the inference time is reduced to 4.7%[32], as shown in Figure 2-3-6.

[32] This data is quoted from the article **"Highly Efficient 8-bit Low Precision Inference of Convolutional Neural Networks with Intel Caffe":** https://arxiv.org/pdf/1805.08691.pdf; the test configuration is as follows: Convolution Model: ResNet50, Hardware: AWS single-socket c5.18xlarge.

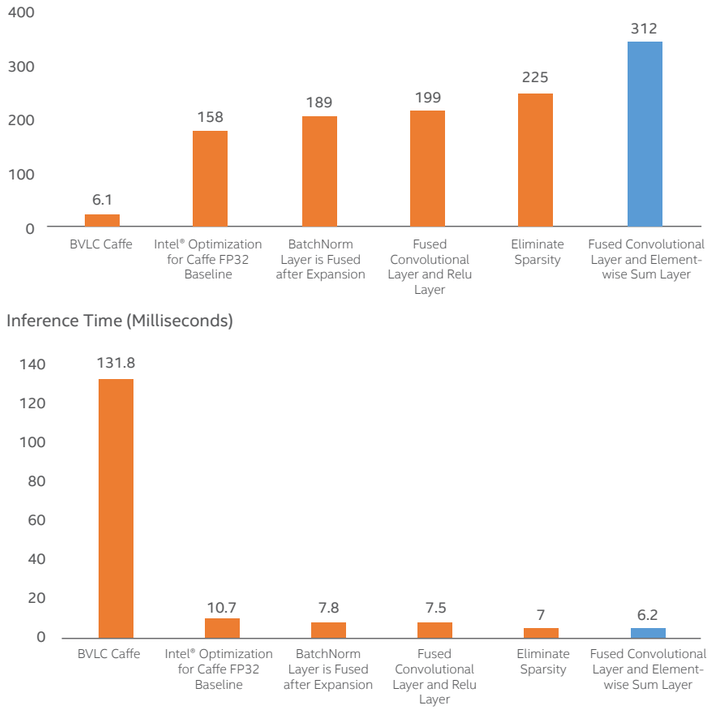


Figure 2-3-6 Intel® Optimization for Caffe Compared with BVLC Caffe in Inference Throughput and Inference Time Performance after being Optimized on Intel® Xeon® Scalable Processor

Intel® DL Boost

Intel® DL Boost, which provides better optimization support for INT8, is incorporated into the new 2nd Generation Intel® Xeon® Scalable Processor. It can accelerate the inference speed of various deep learning models when INT8 is used without affecting the prediction accuracy, and effectively improve the working efficiency of customers' deep learning applications.

In deep learning scenarios such as image classification and object detection, replacing FP32 with lower precision numerical values such as INT8 is a good performance optimization solution. Low precision numerical values can make better use of cache, increase memory data transmission efficiency, reduce bandwidth bottlenecks, and effectively reduce system power while making full use of computing and storage resources. In addition, even when using the same resources, INT8 can achieve higher Operations Per Second (OPS) for deep learning inference.

Intel® DL Boost provides multiple new FMA kernel instructions through VNNI instruction set to support 8-bit or 16-bit low-precision numerical multiplication, which is especially important for deep learning computing that require a large number of matrix multiplications. Intel® DL Boost can also reduce user's requirement for system memory by up to 75%[33] when performing INT8 inference. The reduction of required memory and bandwidth also speeds up the low-precision numerical calculation, thus greatly improving the overall performance of the system.

* For more technical details of about Intel® Xeon® Scalable Processor and Intel® DL Boost, please refer to relevant content in the Technologies section of this guide.

Method for Optimizing Model Accuracy Using Tools

■ Similarity Measurement Tool

In deep learning, the similarity measurement tools can be used to judge the similarity between two features. Different tools can be used for similarity measurement from different dimensions, including the following frequently used ones:

- **Euclidean distance:** It is the most frequently used distance measurement tool. The absolute distance between two points is calculated by such two points in the coordinate system. The greater the distance, the lower the similarity.
- **Cosine similarity in vector space:** The cosine of the angle between two vectors in vector space is used to measure the difference between the two vectors.Compared with distance measurement, cosine similarity pays more attention to the difference of two vectors in direction. The smaller the angle, the higher the similarity.
- **Standardized Euclidean distance:** It is an improved version of Euclidean distance. Before calculating the distance of each feature, it's necessary to standardize the calculation of each component.
- **Mahalanobis distance:** It is used to indicate the distance between a point and a profile. In simple terms, a single sample belongs to the sample set closest to it.

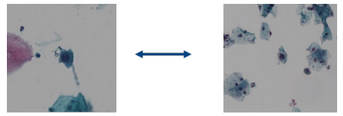


Figure 2-3-7 Using Similarity Measurement Tool to Analyze Causes of Prediction Failure

Similarity measurement tools allow to flexibly design and combine a series of methods to improve the training accuracy of the model. For example, it can be used to analyze the causes of prediction failure by calculating the Euclidean distance between the two features. As shown in Figure 2-3-7, by measuring which positive label the false positive sample is closest to in the feature extraction layer, the main cause of misjudgment can be deduced.

■ Layer-wise Relevance Propagation Tool

Traditionally, the information passing and logic among the layers of the deep learning model have been as difficult to trace back as a black box. The use of Layer-wise Relevance Propagation (LRP) tool can facilitate users solve this problem to a certain extent. LRP tool uses computational relevance to spread the relevance backward layer by layer, showing good traceability. Meanwhile, with this mechanism, the system can also deduce which factors play a greater role in the prediction results, thus improving the accuracy of the model.

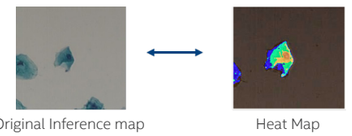


Figure 2-3-8 Using LRP to Detect the Effect of Different Pixels on Inference Results

As shown in Figure 2-3-8, in AI application of medical image analysis and prediction, the LRP tool can be used to see the effect of different pixels on the inference results, and a heat map is formed, thus facilitating the solution to deduce which pixel plays a greater role in the final prediction results.

KFBIO Utilizes AI to Improve Cervical Cancer Screening Efficiency

Background

Cervical cancer is currently one of the malignant tumors that seriously endanger women's health. According to statistics, cervical cancer is the fourth most frequent cancer in women with an estimated 570,000 new cases in 2018 representing 6.6% of all female cancers[34]. However, as the only cancer whose nosogenesis can be known, cervical cancer can be identified early and effectively prevented. Liquid-Based Cytologic Preparation (LBP) screening is easy to operate with high accuracy, which can effectively detect early cancer lesions, facilitate early diagnosis and timely treatment and prevent further spread of cancer cells.

Now, China produces tens of millions of new cervical LBP smears every year, which poses a great challenge to the pathological analysis ability of medical institutions. Therefore, KFBIO, together with Intel, starts to use advanced AI technology to construct and optimize the AI solution for cervical cancer screening based on cervical LBP biopsies, and is committed to promoting effective prevention and treatment of cervical cancer.

At present, there are several factors that pose restrictions on the screening efficiency and accuracy of the solution, making it unable to further improve. The first is data labeling: Compared with other medical data, the analysis data of pathological section has its own unique features. As shown in Figure 2-3-9, pathological section images have a scale ranging from 1 to 40. When scaling at small size, the image cannot be labeled. When the image is scaled at 20 or even 40, only a small part of the whole image can be manually labeled, making it unable to cover all problematic cells in the section.

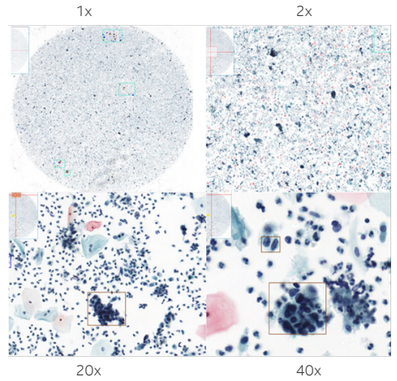


Figure 2-3-9 Pathological Sections in Different Sizes

[33] The data is published on https://software.intel.com/en-us/articles/lower-numerical-precision-deep-learning-inference-and-training

[34] Data published on WTO official website: http://www.who.int/cancer/prevention/diagnosis-screening/cervical-cancer/en/

In addition, in the labeling process, there is also the problem of incomplete labeling. Sometimes, the labeling staff only labels the most serious problematic cells in the field of vision. As shown in the top side of Figure 2-3-10, the malignant tumor in the blue box in the bottom right corner is labeled, but the weak positive cells in the red box in the top left corner are not labeled. On the bottom side of Figure 2-3-10, the labeling position is not accurate enough.

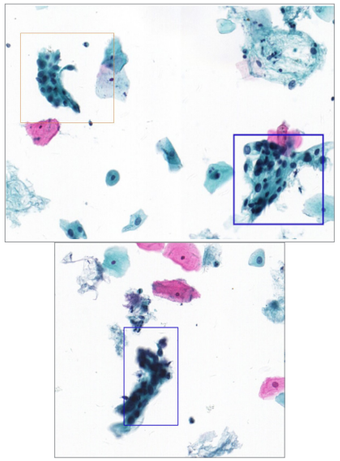


Figure 2-3-10 Pathological Section Image with Incomplete Labeling

Meanwhile, the current labeling scheme usually pays more attention to positive cells and less attention to negative cells. Even if negative cells are labeled, they are only covered at the section level. There is no effective use plan for negative cells that account for the majority of total cells. In addition, the existing labeled samples are seriously unbalanced. Atypical Squamous epithelial Cells (ASC-US) samples account for the vast majority, while Squamous Cell Carcinoma (SCC), endometrium, trichomonas and other samples are few, which hinders the improvement of learning efficiency.

Another issue that needs attention is the selection of neural networks. From the practical results, the frequently used cytopathic object detection network can output the rectangular coordinates of the location of pathological cells and the specific descriptive (The Bethesda System, TBS) classification of pathological cells, but the single object detection network cannot solve the problem of labeling completeness effectively. To solve the above problems, KFBIO, together with Intel, works on optimization from the following dimensions to improve the work efficiency of the screening deep learning model:

- Optimizing data cleaning and pre-processing;
- Constructing a two-stage end-to-end neural network;
- Introducing the model accuracy optimization tool.

Solution and Results

The main workflow of the AI solution for cervical cancer screening based on cervical LBP biopsies built by KFBIO and Intel is shown in Figure 2-3-11. After inputting images, the system obtains prediction of negatives and prediction of positives respectively after going through data pre-processing, classified convolutional neural network and post-processing stages. For prediction of positives, the solution carries out the training of the object detection network (based on ResNet50) model in the second stage, performs the inference process of identification of positives, and submits it to the doctor for final review.

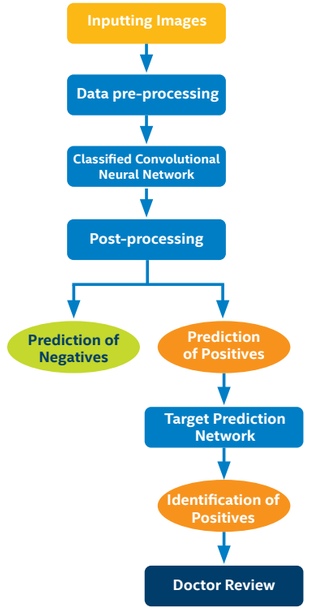


Figure 2-3-11 Optimized Solution Workflow

In the process of optimizing data cleaning and pre-processing, with regard to the problem of different scaling of section images, the solution obtains training data by cutting small images from large section images for pathological section images with larger scaling size and positive labeling as cell/cell mass level. As to the problem of unbalanced samples in the sections, the training set uses the positive/negative ratio of 1:5. Meanwhile, the solution also rotates the samples to expand the diversity of the samples due to the relatively small number of samples with positive labels.

Moreover, in order to improve the utilization efficiency of negative cell samples, the solution assumes that all cells in negative sections are negative cells, and the training set of negative sections is randomly cut from each negative section in proportion (for the purpose of removing section edge interference). Meanwhile, the training set of positive sections is directly cut into 512*512 sub-images according to the coordinate center points marked on the positive sections as well as reasonable random offsets.

To improve the recognition accuracy and efficiency, the solution innovatively builds a two-stage end-to-end neural network. The first stage is classified convolutional neural network and the second stage is object detection neural network. As shown in Figure 2-3-12, the classified convolutional neural network is designed for binary classification inference on the sliding window generated by each section, and fusion of all sliding window results of the section to obtain section-level inference results. The object detection network is used for further positive region detection on the sections determined to be positive in the previous stage.

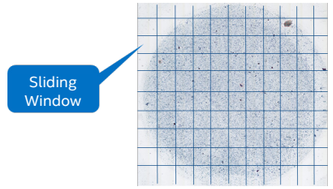


Figure 2-3-12 Classified Convolutional Neural Network Based on Sliding Window Operation

In the process of model training, the solution uses the following optimization scheme to improve the training effect:

- The model uses ResNet50 which has excellent performance on Imagenet dataset for training.
- After the training set is ready, it will be rotated, and then cut to 224*224 based on the center point for normalization and scale processing. Then model training will begin.
- In view of the quite large difference in the number of positive and negative samples in the training set, the solution fuses the sub-images of some negative sections and some positive sections and adds them to the training set incrementally to form iterative training. The positive/negative ratio of the training set is 1:5, further improving the accuracy of the model.
- The solution also adds the similarity measurement tool and the Layer-wise Relevance Propagation(LRP) tool to improve the accuracy of the model.

KFBIO and Intel jointly evaluate the optimized AI solution for cervical cancer screening based on cervical LBP biopsies, conduct training on 5,961 accurately labeled samples, and evaluate different models on 246 test sets.

The evaluation results show that the accuracy of the optimization scheme after adding the classification network has been greatly improved compared with the single object detection network scheme. As shown in Figure 2-3-13, it can be seen that, after adding classification network, when its sensitivity (True Negative Rate, TNR) is 96%, the specificity is close to 70%; in contrast, in the single object detection network scheme, the specificity is only about 40%[35], which means that the accuracy has been greatly improved[36].

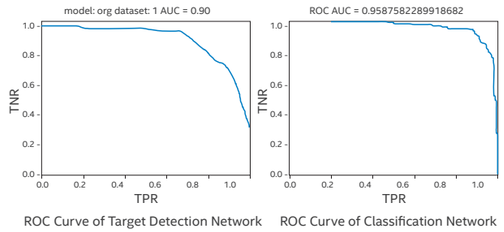


Figure 2-3-13 Comparison of Accuracy between Optimization Scheme and Traditional Scheme

KFBIO Enables Tuberculosis Screening with AI

Background

At present, China is still one of the countries with a high burden of tuberculosis, with about 900,000 new cases of tuberculosis each year, but at the same time, the successful treatment rate of tuberculosis patients in China is more than 90%. An important reason for this contradiction is that the existing tuberculosis screening techniques and methods have yet to be improved. With the rapid development of AI applications in healthcare in recent years, intelligent pathological analysis and diagnosis technology based on deep learning/machine learning methods is gradually being used in tuberculosis screening.

As a biological IT solution provider specializing in the development and production of digital pathology systems, KFBIO is committed to replacing the traditional microscope with a high-precision digital pathology scanner, thereby achieving the digital transformation of traditional pathology sections and using AI-based medical image processing technology to promote intelligent pathology analysis and diagnosis. Now, to address a series of problems limiting the screening and diagnosis of tuberculosis, KFBIO is promoting the application of new intelligent detection technologies with the aid of the fluorescent mycobacterium tuberculosis auxiliary screening system (hereinafter referred to as "Tuberculosis Screening System").

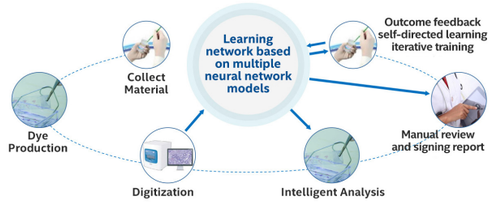


Figure 2-3-14 Basic Workflow of the Tuberculosis Screening System

[35] This data is quoted from **"Pathological Image Analysis Based on Deep Learning published by KFBIO and Intel."**

[36] The data is obtained in the following test configuration: 2S Intel® Xeon® Platinum 8280 Processor, 2.70GHz; Cores/Threads: 28/56; HT: ON; Turbo: ON; Memory: 192GB DDR4 2933; Hard Disk: Intel® SSD SC2KG48; Network Adapter: Intel® Ethernet Network Adapter X722 for 10GBASE-T; BIOS: SE5C620.86B.02.01.0003.020220190234; Operating System: CentOS Linux 7.6; Linux Kernel: 3.10.0-957.el7.x86_64; Compiler Version: ICC 18.0.1 20171018; Caffe Version: Intel® Optimization for Caffe 1.1.0; Workload: ResNet50 with 2 classes, 130 images/second.

Deep Learning-based Tuberculosis Screening System

KFBIO's Tuberculosis Screening System aims to convert the mycobacterium tuberculosis smears into digital images for easy image preservation and transmission, and to develop an auxiliary mycobacterium tuberculosis screening function on the basis of which doctors can greatly improve the interpretation efficiency and solve the problems of objectivity, controllability and repeatability of mycobacterium tuberculosis smear grading.

The basic workflow of the Tuberculosis Screening System is shown in Figure 2-3-14. Thousands of mycobacterium tuberculosis smears will first be scanned using a fluorescence scanner and annotation service platform, and then mycobacterium tuberculosis will be annotated on the scanned documents. This is followed by deep learning based on deep neural networks, allowing the model to accurately identify mycobacterium tuberculosis, as well as the semantic features of the background bacteria/impurities.
To make the system as efficient, reliable and highly available as required for healthcare applications, KFBIO developed the following performance requirements for the system design:

- **Single smear recognition speed:** the recognition necessary for all indicators shall be completed within 180 seconds on general-purpose PC hardware;
- **Mycobacterium tuberculosis detection:** the accuracy for mycobacterium tuberculosis detection AP@[IOU=0.5] > 80%;
- **Quantitative grading of sputum smear negative and positive:** grading accuracy (within 1+) reaches 85% or above.

In order to achieve these goals, KFBIO combined pathology with advanced deep learning/machine learning methods and, as shown in Figure 2-3-15, developed the following technical roadmap:

- In the training phase, after performing a series of steps such as smear scan digitization, data annotation and data enhancement, and foreground detection modeling, a mycobacterium tuberculosis classifier model (e.g. ResNet50) is trained.
- In the application phase, digital images of mycobacterium tuberculosis smears are first acquired using a high-performance digital section scanner, and then the sliding window method is used to extract image patches for deep learning inference. After obtaining inference results for image patches, the Non Maximum Suppression (NMS) algorithm is used to eliminate duplicate objects and objects with low confidence, and finally only high-precision detection results within a single field of view are retained.
- The process of inference and NMS computing is repeated in the above application phase to generate visual results and indicators for full-field recognition, which can be used as input to the auxiliary screening system, thus displaying medical record information, digital images, location/number of mycobacterium tuberculosis and smear grading results for rapid screening and diagnosis.

It can be seen that, compared with traditional computer vision methods, the above new solution based on deep learning methods has advantages such as high detection accuracy, strong morphological adaptability, and more robust models.

Optimized Solution based on Intel Technologies and Its Results

In practice, KFBIO found that existing IT systems in medical organizations are usually built on x86 servers, especially on Intel® servers. To facilitate healthcare organizations maximize the processing performance of their existing IT systems and effectively reduce costs, KFBIO collaborated with Intel to optimize the algorithm model on the Intel® platform and achieved faster inference speed.

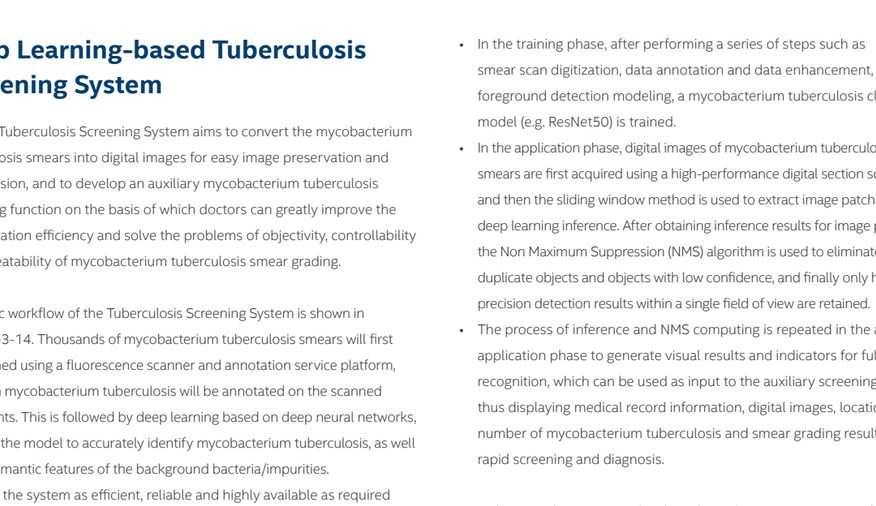


Figure 2-3-15 Technology Roadmap for Auxiliary Mycobacterium Tuberculosis Screening

[37, 38] This data is quoted from internal statistics of KFBIO.

The new optimized solution is built with the profile module that comes with the PyTorch deep model framework, and is applied with the following optimizations after a comprehensive evaluation of modules, kernel running time, processor resource usage, etc.:

- **PyTorch optimization:** The version of PyTorch used before optimization was 1.4, while the new solution upgraded the PyTorch version to 1.6, which optimizes the **native_batch_norm**, resulting in a 22% FPS performance improvement;[37]
- **Memory management optimization:** Considering that the frequent memory request/release process of each framework in the system will consume a lot of resources and time, the new solution introduces jemalloc for dynamic management and memory allocation optimization, which is evaluated to yield about 18% FPS performance improvement;[38]
- **Multi-instance asynchronous processing:** Intel® processors are not only multi-core, but also provide good support for large memory. The new solution adopts multi-instance asynchronous concurrent processing to take full advantage of the multi-core high-memory platform. For example, when using 20 instances for processing, this optimization has been evaluated to achieve about 500% FPS performance improvement;[39]
- **Overall process optimization:** In addition to the above optimizations, the new solution also introduces multi-instance processing, adopts DataLoader, optimizes data input, removes redundant data, etc., so that the final running speed of the system is fully optimized.

In order to evaluate the performance of the optimized solution in actual deployment, KFBIO, together with Intel, evaluated the optimized solution, and the results are shown in Figure 2-3-16. The performance of the fully optimized solution is 11.4 times better than that of the unoptimized solution.[40]

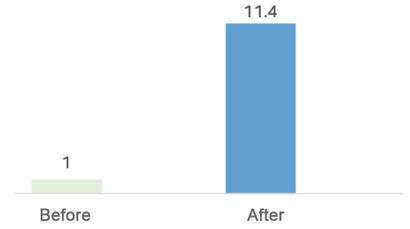


Figure 2-3-16 Comparison of Normalized Performance before and after Solution Optimization

Thanks to the superior performance of Intel® processors and specific optimizations, KFBIO's Tuberculosis Screening System has been widely deployed in many medical organizations. Feedback from field data showed that the new solution was able to maintain 86.8% AP accuracy and 88.9% smear grading accuracy[41] and meet the need to complete digital scanning and quantitative smear grading of a single mycobacterium tuberculosis smear in 80 seconds[42], which was well received by hospitals, doctors and patients alike.

[39] This data is quoted from internal statistics of KFBIO.
[40] Test Workload: Medical Image detection, detectron2 (detectron2 0.1.1), Platform: Dell PowerEdge R740; Processor: 2S Intel® Xeon® Gold 6252 Processor, 2.10GHz; Cores/Threads: 24/48; HT On; Turbo On; Memory 192GB DDR4 (12 x 16384 MB 2666 MT/s); Storage: 1x Intel® 1.8T SSD (Intel® SSDSC2KB01); Network Adapter: Intel® C621 (1 x Intel® X722 for 10GBASE-T); Operating System: Ubuntu 18.04.4 LTS (Kernel: 5.3.0-51-generic); Deep Learning Framework: PyTorch 1.4; Library: Intel® MKL-DNN v0.21.1; Number of Instances: 1; Optimized Solution: Processor: 2S Intel® Xeon® Gold 6252 Processor, 2.10GHz; Cores/Threads: 24/48; HT On; Turbo On; Memory 192GB DDR4 (12 x 16384 MB 2666 MT/s); Storage: 1x Intel® 1.8T SSD (Intel® SSDSC2KB01); Network Adapter: Intel® C621 (1 x Intel® X722 for 10GBASE-T); Operating System: Ubuntu 18.04.4 LTS (Kernel: 5.3.0-51-generic); Deep Learning Framework: PyTorch 1.6; Library: Intel® MKL-DNN DNNL v1.2.0; Number of Instances: 24.
[41] This data is quoted from internal statistics of KFBIO.
[42] Workstation Configuration: Motherboard: X11DPI-N, CPU: Intel Xeon 6240R (24 Core, 2.4GHZ), Memory: 192GB DDR4 (12 x 16GB, 2666MT/S), RAID Card: LSI 9361-8I, Storage: 2x Intel 960G SSD, 4x 4T SATA 3.5 Inch

Conclusion

The use of deep learning method to make rapid detection of pathological section images can not only greatly improve the productivity of pathological examination of medical institutions, eliminate a series of problems caused by the shortage of professional pathologists, but also bring more accurate and timely treatment schemes to patients. Now, the AI application of pathological section detection based on image classification and object detection has received good feedback after being deployed in many medical institutions.

A series of advanced Intel products and technologies, including the Intel® architecture processor platform, Intel® Optimization for Caffe, and Intel® Deep Learning Boost, have greatly improved the work efficiency of pathological section detection application based on deep learning in many application scenarios. For example, the good support of Intel® architecture processor platform for large memory makes it possible to set a larger Batch Size in model training, thus greatly improving training efficiency. Besides, the good support of Intel® Optimization for Caffe and Intel® Deep Learning Boost for INT8 can effectively improve the inference efficiency and real-time analysis of pathological sections.

Although the processor platform involved in this case is the 1st Generation Intel® Xeon® Scalable Processor, with the arrival of the new 2nd Generation Intel® Xeon® Scalable Processor and other new Intel products and technologies, users can build AI applications with more powerful training and inference performance based on these updated software and hardware. Meanwhile, Intel also plans to conduct inference optimization research on more deep learning models to help more patients get valuable treatment time and efficiency.

AI Technology Assists in Drug Research and Development

Deep Learning Accelerating Drug Screening

Phenotypic Classification Based on HCS

More and more new technologies are being applied to accelerate the drug research and development process. The method of High Content Screening (HCS) based on cell images is one of the frequently used automated analysis methods in the field of system biology and drug research and development, and is also an important application of AI technology in the early stage of drug discovery. It analyzes and obtains cell phenotype features induced by genetic or chemical treatment through image information obtained by microscopic imaging techniques.

In this process, phenotypic detection, analysis and classification of cell images matter the most. However, the inherent complexity of biological analysis process and the inherent variability of cell measurement have brought severe challenges to the analysis of phenotypes in cell images. The traditional image analysis method for cellular phenotypic feature extraction mainly consists of a series of independent data analysis steps. As shown in Figure 2-4-1, after the original image is input, object detection method is first used for feature extraction at cell level or image level. Then, these features are converted (selected, standardized, etc.). Finally, relevant features are summarized, and used as input to the classification algorithm for phenotype prediction.

Although the above feature detection, analysis and classification methods have been successfully applied in a large number of drug research and development processes, they are still subject to limitations. For example, for object segmentation, dimensionality reduction and phenotypic classification, a large amount of priori knowledge is usually required (for example, the geometric properties of the expected phenotypes) to customize each measurement process. At the same time, the performing of each step using traditional HCS method involves method customization and parameter adjustment. However, during the performance optimization of the whole analysis process, there are still many challenges as to how to jointly optimize all parameters to achieve the best performance, so the overall efficiency still needs to be improved. Therefore, more AI methods based on deep learning are gradually being introduced into HCS phenotype classification based on cell images.

DL-based HCS Method[43]

Background

In the traditional HCS image analysis method, image data is converted to different levels of abstraction, such as pixel intensity, and so on. In deep learning approaches such as deep neural networks, hierarchical abstraction in these image data can be calculated and analyzed through a framework, but these methods rely heavily on manually defined features. In contrast, CNN can automatically learn and extract features from images, so it has better efficiency in phenotypic prediction of cell images.

CNN network usually includes input layer, convolutional layer, ReLU layer, pooling layer, fully connected layer, and other layers. The convolutional layer obtains two-dimensional geometric information in an image by calculating convolution between layer input (e.g., the output of an original image or a previous convolutional layer) and multiple two-dimensional convolutional kernels. Each convolutional kernel can encode a geometric pattern, and can be convolved to obtain a convolutional kernel map (or feature map), which is a pixel-based non-linear activation function and will be transferred to the subsequent convolutional layer for a more complex pattern. Finally, the output of the convolutional layer is transfered to the fully connected layer, and a prediction is generated for a given input in a feed-forward way.

Assuming that the output layer of CNN has $\mathbf{N_p}$ phenotypes to be classified, then for a given input image $\mathbf{x}$, the network will calculate the activation function $\mathbf{a_j(x)}$ of $\mathbf{j}$ unit in each socket at the output layer, and calculate a vector $\boldsymbol{\rho}$ based on this. $\boldsymbol{\rho_k}$ can form a probability quality function for covering $\mathbf{N_p}$ phenotypes to be classified:

$$\rho_k := p(y = k|\mathbf{x}) = \frac{\exp(a_k(\mathbf{x}))}{\sum_j^{N_p} \exp(a_j(\mathbf{x}))}$$

Where, $\mathbf{k}$ is the serial number of the phenotype, and according to these probabilities, the following predicted value of the phenotype can be obtained:

$$\hat{y} = \underset{k}{\operatorname{argmax}}\ p(y = k|\mathbf{x})$$

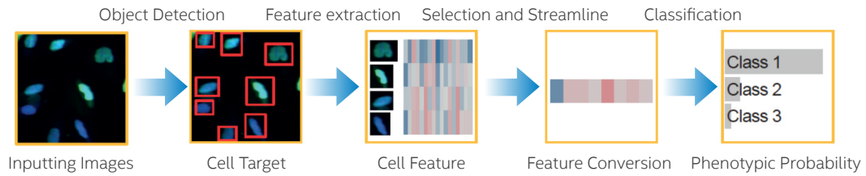


Figure 2-4-1 Traditional HCS Method

[43] For the technical description of HCS based on CNN and M-CNN in this section, please refer to: Godinez et al, A multi-scale convolutional neural network for phenotyping high-content cellular images. Bioinformatics, 2017

Therefore, it can be seen that such factors as the number of layers, the number of units in the convolutional layer, and the selection of convolutional kernel and pooling factor will all affect the prediction performance. However, in cell phenotype classification, there is another problem, that is, due to different cell sizes and microscopic imaging sizes, there are often large spatial differences in image data, which may lead to a decline in accuracy if the classical CNN network structure continues to be used.

Multi-scale Convolutional Neural Networks (M-CNN) can better solve this problem. Compared with the classical CNN network structure, it adds parallel multiscale analysis. For images at different scales, different CNN networks can be used for training in an independent way.

Figure 2-4-2 shows a M-CNN network structure with 7 scales, and the scaling size changes from top to bottom. The network will input the image at seven different scales at its input layer, and use the sequence of three convolutional layers to process scaled images at each scale. The convolutional path of each scale is independent of other scales, and at the last layer of each scale, the convolutional kernel map obtained is scaled to the largest size with the aggregation method and linked together to be used as the input of the final convolutional layer, and the final output layer will output the generation probability value of each phenotype.

Recommended Hardware and Software Configuration

For accelerating drug research and development by using AI technology, refer to the following hardware and software configuration based on Intel® architecture platform for system deployment.

Name	Specification
Processor	Intel® Xeon® Gold 6240 processor or higher
Hyper Threading	On
Tutbo Boost	On
Memory	16GB DDR4 2666MHz* 12 and above
Storage	Intel® SSD D5-P4320 Series and above
Operating System	CentOS Linux 7.6 or latest version
Linux Kernel	3.10.0 or latest version
Compiler	GCC 4.8.5 or latest version
TensorFlow Version	Intel® Optimization for TensorFlow v1.7.0 or latest version
Horovod	0.12.1 or latest version
OpenMPI	3.0.0 or latest version
ToRSwitch	Intel® Omni-Path Architecture

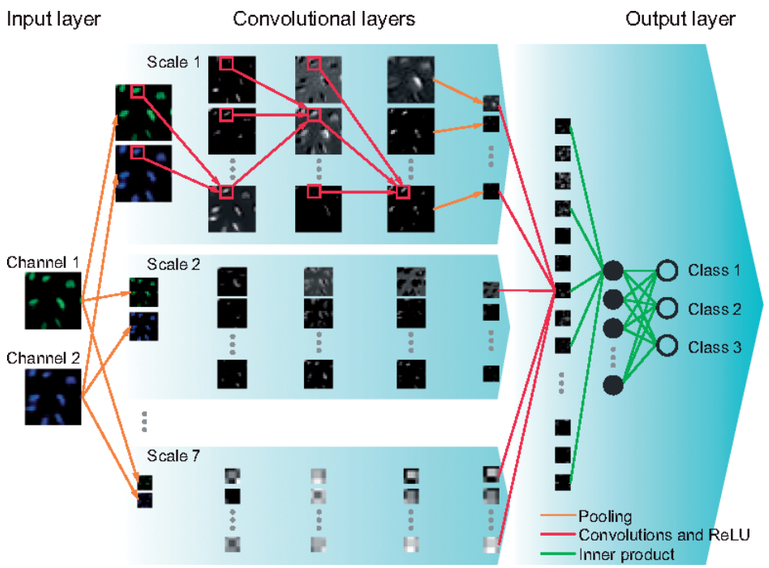


Figure 2-4-2 M-CNN Architecture Diagram

Optimizations Powered by Intel® Xeon® Scalable Processor

Improving the Training Efficiency of Single Computing Node

It often takes several years for the research and development of a new drug, and what comes with it is the anxious waiting of patients. To further improve the efficiency of method of HCS based on M-CNN network model in drug discovery and further accelerate research and development, a series of optimization schemes for Intel® Xeon® Scalable Processor have been introduced, which include various methods such as increasing throughput of single computing node and increasing efficiency of multiple computing nodes.

Firstly, the code for starting the M-CNN model on a single computing node for training is as follows:

```
python tf_cnn_benchmarks.py
--model=mcnn
--batch_size=32
--data_format=NCHW
--data_dir=INPUT_DATA_DIR
--data_name=mcnn
--num_intra_threads=40
--num_inter_threads=2
--num_batches=2000
--num_warmup_batches=70
--display_every=5
--momentum=0.9
--weight_decay=0.00005
--optimizer=momentum
--resize_method=bilinear
--distortions=False
--sync_on_finish=True
--device=cpu
--mkl=True
--kmp_affinity="granularity=fine,compact,1,0"
--variable_update=horovod
--local_parameter_device=cpu
--kmp_blocktime=1
--train_dir=TRAIN_DATAWRITE_DIR
```

On a single computing node, one of the problems encountered by M-CNN method is the memory capacity. Generally speaking, the efficiency of deep learning network can be improved to a certain extent with the increase of Batch Size. Cell images used in high content screening usually have a large size, and in multi-scale joint operation, when the Batch Size is increased to a certain amount, quite a large amount of memory capacity will be required. As shown in Figure 2-4-3, when the Batch Size is 32, the system requires 47.5GB of memory.

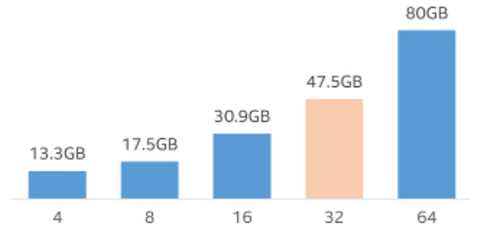


Figure 2-4-3 Memory Demand under Different Batch Sizes

The Intel® Xeon® Scalable Processor platform has good support capability for large memory, which can effectively meet the large memory demand brought by the increase of Batch Size. Its more optimized microarchitecture, more cores, and control and scheduling capability for faster and larger memory capacity make the method of M-CNN based on TensorFlow framework easy to deploy. In a test using the Broad Bioimage Benchmark Collection 021 (BBBC-021) dataset [44], the input microscope image size is 1024*1280*3. When the Batch Size is 32, the processing speed reaches 13 images per second under a single TensorFlow worker. However, such processing speed seems to be slow for datasets with thousands of images. The whole training process is still very long and the efficiency needs to be improved.

With the introduction of NUMA technology and the weight synchronization technology based on distributed deep learning framework Horovod, users can simultaneously use four TensorFlow workers under TensorFlow framework. As shown in Figure 2-4-4, a two-socket Intel® Xeon® Scalable Processor deployed in a typical computing node can be divided into four computing areas, each of which executes a TensorFlow worker.

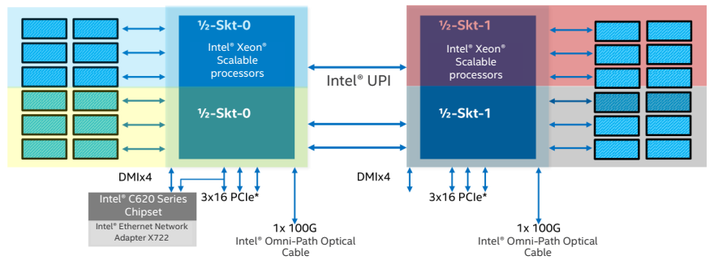


Figure 2-4-4 Division of Two-socket Intel® Xeon® Scalable Processors in Typical Computing Nodes

[44] BBBC-021: Ljosa V, Sokolnicki KL, Carpenter AE, Annotated high-throughput microscopy image sets for validation, Nature Methods, 2012

The technical characteristics of NUMA allow for binding different cores and different memories of processors to perform training without competing over computing resources and storage resources. Intel® Ultra Path Interconnect (Intel® UPI) allows for weight synchronization across computing areas. In this way, the throughput of the training model can be further improved. As shown in Figure 2-4-5, after using four TensorFlow workers, when the Batch Size is still 32, the processing speed reaches 16.3 images per second, and the efficiency increases by 25.4%.

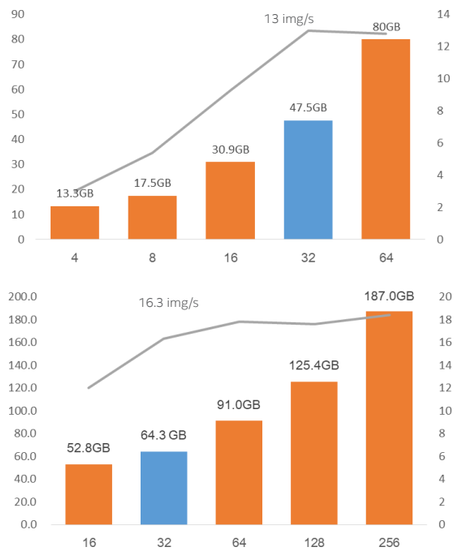


Figure 2-4-5 Performance Comparison of Four Worker Threads and Single Worker Thread in TensorFlow

Improving the Training Efficiency of Multiple Computing Nodes

In addition to improving the training efficiency of single computing node, the distributed training technology can also further improve the training efficiency of multiple computing nodes. In the classical TensorFlow distributed architecture, the parameter server is needed to average the gradient, and each processing thread may be used as a worker thread or parameter server. The former is used for user processing and training data, calculating gradients, and transferring them to the parameter server for averaging.

However, in this method, if the processing capacity of the parameter server is insufficient, it may cause an overall bottleneck of the system. Meanwhile, in order to achieve optimal performance, users need to specify appropriate initial worker threads and parameter servers at the beginning, but a slight carelessness will lead to performance decline. Horovod, a new open source distributed deep learning framework for TensorFlow, can effectively solve this problem. The Ring-allreduce algorithm introduced in it builds a new communication strategy, allowing the worker threads to average the gradient without adding any parameter server.

As shown in Figure 2-4-6, in Ring-allreduce algorithm, each worker thread first performs gradient calculation according to its own training data to obtain gradient information. Each worker thread communicates 2*(N-1) times with other N-1 workers. In this process, a worker thread sends and receives gradient information from the data buffer. The gradient information received each time is added to the worker process buffer to replace the previous value. After sending and receiving N-1 gradient messages, all worker threads will receive the gradients required to calculate and update the model. This method can maximize the use of network capacity to avoid calculation bottleneck [45]. Based on this communication strategy, Horovod establishes a distributed system based on TensorFlow with the Open Message Passing Interface (OpenMPI).

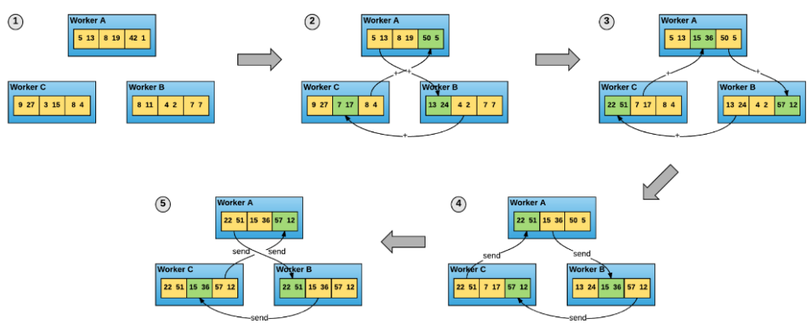


Figure 2-4-6 Ring-allreduce Algorithm Diagram

Even under the Horovod framework, there are still a considerable amount of gradient information needed to be transferred. For example, in the test using the BBBC-021 dataset, the gradient information size is 162.2MB.

Intel® Omni-Path architecture supported by Intel® Xeon® Scalable Processor makes gradient information transfer faster, thus improving the overall training efficiency of M-CNN method. Intel® Omni-Path architecture has 100Gbps point-to-point bandwidth and 1us-level point-to-point MPI communication latency. It is fully compatible with OFA software interface, fully supports RDMA and PSM interfaces, and utilizes innovative technologies such as message packet integrity protection and dynamic link expansion, which lays a solid foundation for high-speed transmission of gradient information. As shown in Figure 2-4-7, in 8 nodes deployed with Intel® Xeon® Scalable Processor, the synchronization point transmission is greater than 10Gb under Horovod framework.

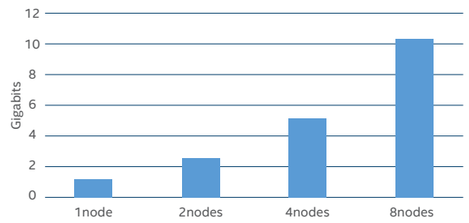


Figure 2-4-7 Synchronization Point Transmission Using Horovod and Intel® Omni-Path Architecture is Greater than 10Gb

Another way to optimize the training efficiency of multiple computing nodes is to converge and adjust the Learning Rate (LR). LR size in different training stages is a very important setting item in deep learning. If LR is too large, it will cause oscillation, and if LR is too small, it will cause slow convergence speed and easily result in over-fitting. In the training process of M-CNN model based on TensorFlow framework, the following LR adjustment methods can be used for performance optimization.

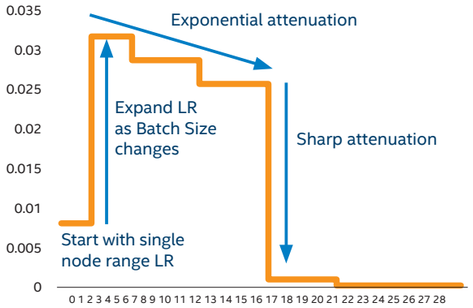


Figure 2-4-8 LR Adjustment during M-CNN Network Training

As shown in Figure 2-4-8, at the beginning of training, the first iteration uses a single node LR, and then expands it to the global Batch Size parameter. In the following iterations, LR attenuates exponentially, and attenuates sharply from the 14th iteration [46].

Thus, the training command of the M-CNN network on the multiple computing nodes is as follows:

```
OMP_NUM_THREADS=10 mpirun -np 32 -cpus-per-proc 10
--map-by socket -hostfile HOSTFILE
--report-bindings
--oversubscribe -x LD_LIBRARY_PATH -x
PATH -x OMP_NUM_THREADS -x HOROVOD_FUSION_THRESHOLD numactl -l

python tf_cnn_benchmarks.py
--model=mcnn
--batch_size=8
--data_format=NCHW
--data_dir=INPUT_DATA_DIR
--data_name=mcnn
--num_intra_threads=10
--num_inter_threads=2
--num_batches=2000
--num_warmup_batches=70
--display_every=5
--momentum=0.9
--weight_decay=0.00005
--optimizer=momentum
--resize_method=bilinear
--distortions=False
--sync_on_finish=True
--device=cpu
--mkl=True
--kmp_affinity=="granularity=fine,compact,1,0"
--variable_update=horovod
--local_parameter_device=cpu
--kmp_blocktime=1
--horovod_device=cpu
--piecewise_learning_rate_schedule='0.008;2;0.032;5;0.029;10;0.026;15;0.001;20;0.0001'
--train_dir=TRAIN_DATAWRITE_DIR
--save_summaries_steps=1
--summary_verbosity=1
```

[45] For detailed technical description, please refer to: Alex Sergeev, Mike Del Balso, Meet Horovod: Uber's Open Source Distributed Deep Learning Framework

[46] For technical details about LR setting, please refer to: Yang You et el, 2017, "ImageNet Training in Minutes"

Novartis Utilizes Deep Learning to Improve Drug R&D Efficiency

Background

As a leading pharmaceutical enterprise in the world, Novartis is actively taking advantage of digital transformation to maintain its competitive advantages in drug innovation, disease diagnosis and drug research, and "AI + Drug Discovery" is an important link in its future drug research and development process.

Now, Novartis is working with Intel to study how to speed up the HCS process using deep learning. Cell phenotype in HCS is one of the important methods used by Novartis for drug discovery in early stage. The so-called high content refers to a rich set of thousands of predefined features (e.g. size, shape, texture.) extracted from images using classical image processing techniques. HCS allows microscopic images to be analyzed to study the effects of thousands of genetic or chemical treatments on different cell cultures. With the deep learning approach, Novartis can "automatically" learn from the data and distinguish the relevant image features of one treatment from another. However, it still requires a considerable amount of time for this method due to the huge amount of information in cell microscope images - the training time of its image analysis model is about 11 hours[47].

Now, biologists and data scientists from Intel and Novartis hope to accelerate HCS analysis with the M-CNN network deployed on the optimized Intel® Xeon® Scalable Processor platform. In this joint effort, the team focuses on the entire microscope image instead of first identifying each cell in the image in separate processes. Moreover, the microscope images in the BBBC-021 dataset it uses may be much larger than the images in common deep learning datasets.

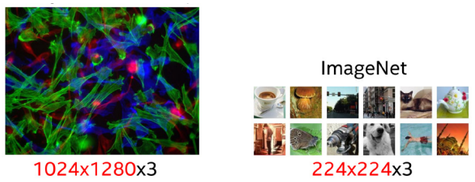


Figure 2-4-9 Comparison of Microscope Images for HCS with Common Image Dataset

As shown in Figure 2-4-9, the left is a microscope image for HCS, with a single pixel close to 4,000,000 while the right is an image from the famous ImageNet dataset[48], with a single image of 150,000 pixels for its training dataset, a 26-fold difference between the two. Large-scale microscope images together with the millions of parameters brought by them, plus the thousands of training images at a time, not only pose challenges to the system memory, but also bring a huge computational load. In order to effectively deal with this challenge, the two sides utilize a series of deep neural network optimization and acceleration technologies to facilitate the system process multiple images in a shorter time while maintaining accuracy.

Optimized Solution and Results

The optimization scheme accelerates the training of M-CNN model deployed on Intel® Xeon® Scalable Processor platform in the following two aspects. Firstly, in the single computing node, it takes full advantage of the good support of Intel® Xeon® Scalable Processor platform for large memory, so that it can use large Batch Size (set to 32 in the scheme), and it utilizes NUMA technology to increase worker threads to improve training efficiency. Secondly, in the multiple computing node, it introduces Horovod, an open source distributed deep learning framework for TensorFlow, and combines it with Intel® Omni-Path architecture supported by Intel® Xeon® Scalable Processor to greatly improve the training efficiency of M-CNN model under multiple nodes. Meanwhile, it designs and adopts the optimized learning rate convergence and adjustment method to improve the performance[49].

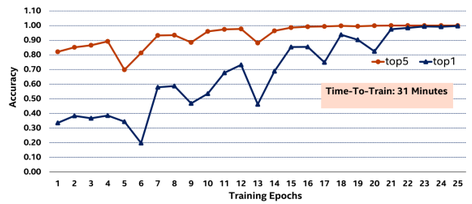


Figure 2-4-10 Training Effect of Novartis's Optimized Scheme

The scheme deploys eight nodes based on Intel® Xeon® Scalable Processor, and uses the BBBC-021 dataset, with a total of 10,000 images whose size is 1024*1280*3. After more than 20 times of training, as shown in Figure 2-4-10, the total training time is about 31 minutes, with the accuracy exceeding 99%. Meanwhile, after using NUMA technology to form 32 TensorFlow workers (4 worker threads per node), the processing capacity reaches more than 120 images per second for the scheme, and the performance is significantly improved compared with that before optimization.

Conclusion

It takes at least several years for a new drug to be discovered, tested and produced, and what comes with this is the ardent expectations from patients and their family members. Using AI technology to speed up the process of drug research and development is not only a common choice for many pharmaceutical enterprises to accelerate innovation and maintain their core competitiveness, but also an important way to enable science and technology to manifest itself in benefiting mankind and creating a healthy life. To this end, Intel is also working with many pharmaceutical enterprises to accelerate the application of AI solutions in drug research and development.

Through reasonable optimization schemes, advanced technologies and products such as Intel® Xeon® Scalable Processor and Intel® Omni-Path architecture can provide excellent and reliable large memory support for AI applications such as DL-based HCS, as well as large Batch Size and multiple TensorFlow workers support to improve the training efficiency of single node or multiple nodes, and provide support for Horovod distributed training framework with high bandwidth and low latency, thus greatly accelerating the drug research and development process of pharmaceutical enterprises such as Novartis.

Now, a series of AI applications based on Intel® Xeon® Scalable Processor platform have been deployed in many pharmaceutical enterprises, which have produced good results. It is worth mentioning that although the test in this guidebook is based on the Intel® Xeon® gold 6148 processor platform, with the introduction and application of the next generation Intel hardware and technologies such as the 2nd Generation Intel® Xeon® Scalable Processor and Intel® Optane® persistent memory, users can choose the updated Intel hardware platform and related software optimization schemes in future actual deployment to build a deep learning scheme with stronger performance, and obtain better training and inference effects, thus further accelerating the process of drug discovery and better assisting patients in treatment and rehabilitation.

[47] This data is quoted from: https://newsroom.intel.com/news/using-deep-neural-network-acceleration-image-analysis-drug-discovery/#gs.ptk50k
[48] ImageNet: Russakovsky O et al, ImageNet Large Scale Visual Recognition Challenge, IJCV, 2015
[49] The data is obtained in the following test configuration: 2S Intel® Xeon® Gold 6148 Processor, 2.40GHz; Cores/Threads: 20/40; HT: ON; Turbo: ON; Memory: 16GB DDR4 2666*12; Hard Disk: 480GB Intel® SSD OS drive*1, 1.6TB Intel® SSD data drive*1; Network Adapter: Intel® Omni-Path Host Fabric Interface (HFI); BIOS: SE5C620.8 6B.02.01.0008.031920191559; Operating System: CentOS Linux 7.3; gcc Version: 6.2; Tensorflow Version: Intel® Optimization for TensorFlow v1.7.0; Horovod Version: 0.12.1; OpenMPI: 3.0.0; ToRSwitch: Intel® Omni-Path architecture workload: Broad Bioimage Benchmark Collection* 021 (BBBC-021) dataset, 10,000 images with a size of 1024*1280*3.

Machine Learning Helps Create More Accurate and Intelligent Healthcare Solutions

Machine Learning Methods Used in the Healthcare Industry and their Trends

More machine learning methods are being applied in the healthcare industry

As an important approach to achieve AI, traditional machine learning has also been widely used in the healthcare industry. Unlike the above mentioned deep learning methods, the applications of machine learning in healthcare are more focused on health status insights, auxiliary disease diagnosis and treatment, image feature extraction, and pathology research by relying on massive amounts of healthcare data.

With the development of medical informatization, medical data has gradually moved from paper records to electronic records in the past few decades, which provides the data basis for the application of machine learning. Many healthcare organizations have already begun to deploy various machine learning-based AI technologies and achieved good results in medical research and clinical support.

For example, some healthcare organizations are trying to use decision trees, Random Forest (RF), and other machine learning algorithms to analyze massive amounts of data on diabetic patients and predict the probability of diabetes. Data comparison shows that, machine learning-based methods for predicting the probability of diabetes are more efficient than human expertise. In other healthcare organizations, machine learning models built on massive data are helping doctors efficiently assess the prognostic risk scores for patients with coronary heart disease, in order to better determine their clinical prognosis and choose the best treatment plans.

In addition to aiding in the diagnosis and treatment of known diseases, machine learning methods can also facilitate healthcare organizations comb through large volumes of complex medical records to predict unknown disease signals, such as prediction of cardiovascular disease risk factors and refractive errors from retinal fundus images. This article will then describe several important trends of machine learning application in healthcare, including precision epidemic prevention and control, chronic disease prevention and treatment, and radiomics applications, and introduce relevant practical cases to discuss the development trend of machine learning in healthcare.

Important trends of machine learning applications in healthcare

■ Precision Epidemic Prevention and Control

The unexpected explosion of COVID-19 epidemic has brought renewed attention to epidemic prevention and control. As an important branch in healthcare, epidemic prevention and control requires the concerted efforts from all walks of life to stop the spread of the virus and end the epidemic through propaganda, protection, isolation and tracing. In addition, efficient testing methods are required to quickly screen infected people, so as to effectively arrange medical resources and provide better treatment plans. At the same time, precision epidemic prevention and control also requires an understanding of how and where the virus spreads, in order to trace back to the source and accelerate the vaccine and drug R&D.

Traditional epidemic prevention and control cannot be achieved without encouraging and mobilizing "human resources", such as mobilizing community volunteers and primary health care workers to conduct follow-up visits and temperature checks, and then reporting, summarizing and analyzing the information obtained. Today, with the rapid development and popularization of new technologies such as the Internet, big data, cloud computing and artificial intelligence, new methods of epidemic prevention and control, such as AI technology, are gradually playing an important role in epidemic surveillance and analysis, virus tracing, patient tracking, personnel movement, and community management.

Although AI technology plays an increasingly important role in epidemic prevention and control, it is worth noting that the large amount of data obtained in epidemic prevention and control are often discrete data, which is nearly impossible to annotate on a large scale, and even if there are some supervised data, they may be a small dataset that contains dozens or hundreds of entries. Therefore, the machine learning approach is more technically advantageous in this case. As the advanced hardware infrastructures such as Intel® processors provide greater computing power for a variety of machine learning algorithms, the machine learning-based epidemic prevention and control methods are playing an increasingly significant role in pathology detection, precision prevention and control, epidemic simulation, and transmission path analysis.

One of the key measures to prevent and control the epidemic is to enable governments, epidemic prevention and control agencies, healthcare organizations and research institutions at all levels to quickly understand the development trend of the epidemic, introduce reasonable and effective control methods, implement precise epidemic prevention and control, and accelerate the research and development of vaccines and antiviral drugs. As mentioned before, machine learning methods such as classification, regression, and clustering can be used to build a comprehensive epidemic prevention and control solution, as shown in Figure 2-5-1.

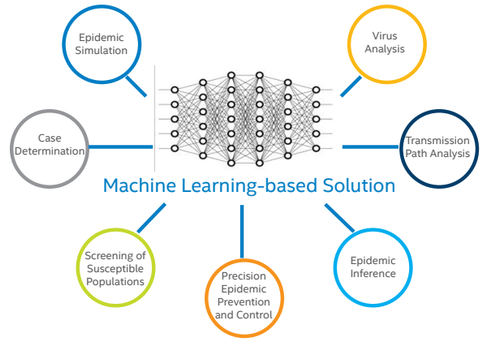


Figure 2-5-1 Machine Learning-based Epidemic Prevention and Control Solution

Firstly, it is extremely critical to analyze the path of virus transmission following an outbreak. Machine learning methods are used to build the transmission path model that simulates a network of potential transmission relationships, which can quickly identify possible transmission paths to aid in precise prevention and control. Secondly, at the critical stage of prevention and control, the key is to identify potential high-risk populations. The machine learning classification algorithms can help build more accurate

screening models and use AI to enrich the existing model of prevention and control screening rules, thereby further improving the coverage, recall and accuracy of screening.

Many machine learning algorithms can be used to speed up clinical diagnosis in epidemics. For example, in the clinical diagnosis of COVID-19, the patient's most prevalent clinical manifestations are fever, fatigue, and dry cough, which are similar to those of influenza and other diseases. A machine learning approach to building clinical diagnostic models can facilitate healthcare organizations quickly identify infected patients to ensure the optimal use of healthcare resources.

Finally, the machine learning methods also accelerate the virus analysis. The structure of the COVID-19 virus consists of a protein shell and its internal genetic material (Positive-strand RNA), which is able to use itself as a template to continuously replicate and then infect more normal cells. Thus, an efficient genetic analysis of COVID-19 virus can accelerate the R&D of vaccines and antiviral drugs.

■ Chronic Disease Prevention and Treatment

Along with the lifestyle changes brought about by industrialization and urbanization, as well as the accelerated aging of the population and the impact of unhealthy lifestyles, chronic non-communicable diseases such as cardiovascular and cerebrovascular diseases, diabetes and chronic respiratory diseases (hereinafter referred to as "Chronic Diseases") have become the main cause of death of Chinese residents. A piece of data shows that, since the new century, more than 25% of China's adult population has been suffering from hypertension, and the mortality rate from chronic diseases has accounted for more than 86% of the total mortality rate[50]. Therefore, chronic diseases have gradually become a major public health problem.

Compared to infectious diseases, pathogenic infections, food poisoning and other acute diseases, chronic diseases have the following characteristics:

- Chronic diseases are predominantly prevalent among the elderly, and their prevalence increases with age;
- Chronic diseases are often lifelong, with long-term treatment, care and rehabilitation period, high demand for medical services and high care requirements;
- Most chronic diseases are irreversible, affecting the quality of life of patients and imposing a heavy financial burden on families and society;
- Chronic diseases often have cross-complications, which make it difficult for a single treatment plan to be effective and require comprehensive rehabilitation.

Considering these characteristics, healthcare organizations at all levels develop a strategy of "Prevention is the best cure" for the treatment of chronic diseases, but this requires a comprehensive evaluation on the health status of patients and long-term follow-up visit. However, the existing traditional medical practice that focuses on specialist clinics and annual medical check-ups obviously cannot achieve early screening, early detection and early treatment.

The gradually accumulated rich and diverse medical data lays the foundation for the application of AI in the prevention and treatment of chronic diseases. By utilizing certain algorithms, machine learning can discover patterns in various patient health data and learn these patterns through modeling, thereby predicting chronic diseases.

Taking diabetes prediction as an example, monitoring glycated haemoglobin (HbAlc) has traditionally been recognized as the most reliable "gold standard" for diagnosing diabetes, but this method requires a blood test, which can be affected by medication use, changes in health status, and, more importantly, this clinical test cannot predict the future chances of developing the disease. Now, by using machine learning methods, some research institutions are able to build predictive model of diabetes from hundreds of thousands of records by collecting indicators such as blood glucose, blood pressure, lipids, fasting insulin, BMI, and age. This model not only can determine the current state of the patient, but also predict the likelihood of the patient developing the disease in the next few years, thus facilitating healthcare organizations to develop appropriate preventive measures.

By deploying similar chronic disease prevention and treatment algorithms in healthcare organizations, rehabilitation centers, and even home smart devices, people at high risk for chronic diseases such as the elderly, obese, and smokers can more easily receive chronic disease risk assessments, personalized health interventions, and long-term evaluations of intervention effects, thereby enabling better self-health management.

■ Radiomics Application

Since it was first mentioned in 2012[51], radiomics has been of great interest to the healthcare industry. Radiomics refers to the extraction of a large amount of image information from CT, MRI, PET and other medical images in a (semi-) high-throughput manner, and then performing processes such as area segmentation, feature extraction and model to carry out deeper mining, prediction and analysis of image data, thus assisting doctors in making more accurate diagnoses. Radiomics has played an increasingly critical role in the diagnosis and treatment of diseases such as cancer.

The reason is that, on the one hand, the temporal and spatial heterogeneity of tumor genomes (i.e., tumors show molecular biology or genetic changes at different times in the growth process) limits the effectiveness of targeted therapies, and traditional medical organizations lack effective means to comprehensively and quantitatively evaluate tumor heterogeneity; on the other hand, since tumors must be larger than 5 mm before they can be effectively detected and diagnosed, it is often the case that the cancer is irreversible when it is discovered, resulting in the five-year survival rate of patients in the past maintaining at a low level.

By combining genetic information and image multi-modal information, radiomics can convert images into mineable high-throughput image feature data, quantify the spatial and temporal heterogeneity of tumor tissues, reveal disease features that cannot be identified with the naked eye, effectively convert medical images into a high-dimensional identifiable feature space, and use statistical and/or machine learning methods to screen the most valuable imaging features for clinical analysis, thereby building models that can be used for diagnostic, prognostic, or predictive purposes and provide valuable information for accurate personalized diagnosis and treatment. Compared with the biopsy, radiomics analysis can not only extract disease features comprehensively, but also reuse data; and compared with traditional medical imaging, radiomics can benefit from high throughput, quantitative analysis, fast computing speed, and high precision, thus it has received widespread attention and research from researchers.

The basic analysis workflow of radiomics is shown in Figure 2-5-2, which is divided into several major steps including data acquisition, Volumes Of Interest (VOI) segmentation, feature extraction, feature selection, model training, and model evaluation and prediction. At the data acquisition phase, the system will import the patient's medical images such as CT, MRI, PET-CT, in DICOM image format, and the gene expression profile and clinical report will be loaded into the system in a specific clinical information format, and the imported data should have the same or similar acquisition parameters to ensure that the data will not be affected by the machine model and parameters.

Considering that the image data included in a study may come from scanning machines with different scanning parameters or from different scanning machines, in order to minimize the resulting differences in image data and their impact on the final results, the platform will resample each image and interpolate them by using the BSpline function, so as to ensure that each group of images has the same resolution in post-processing and to normalize the signal.

VOI segmentation is the process of sketching a region of interest on an image image so that radiomics features can be calculated for this specific region. Feature extraction is a comprehensive analysis of tumor heterogeneity by extracting intensity, shape and other features and combining low-dimensional visual features, high-dimensional complex features and clinical experience features.

Then, after filtering specific radiomics features with feature selection methods such as Least Absolute Shrinkage and Selection Operator (LASSO) regression filtering and Principal Component Analysis (PCA), the model is trained by using machine learning methods such as logistics regression and decision trees. Finally, the system evaluates the model results by using Receiver Operating Characteristic Curve (ROC), Nomogram, etc., and makes prognostic predictions.

In the above process, the selection of the appropriate feature selection method will affect the prediction efficiency and accuracy of the entire radiomics system. Usually, according to the different image acquisition parameters, as well as the interference caused by respiratory motion displacement, the system needs to use a reasonable feature selection method to filter radiomics features with high noise immunity and improve their stability by adjusting the parameters. In addition, feature selection is a key step in avoiding "Curse of Dimensionality" and information loss.

To address the lack of IT infrastructure in healthcare organizations, when providing radiomics solutions, IT solution providers are also introducing data visualization tools to help doctors and allow for one-click operation. After inputting a large number of lesion images, the radiomics solution can quickly provide statistically valuable features for machine learning models to train by extracting and normalizing the features with one click, thus effectively improving the efficiency and accuracy of the models.

Nowadays, a series of radiomics-based medical research and auxiliary diagnosis and treatment solutions have been deployed and implemented in many healthcare organizations, and have achieved remarkable results in early cancer screening and other scenarios. It is worth mentioning that radiomics solutions can also be used for the precise recognition and diagnosis of COVID-19. In CT images of COVID-19 patients, the lesions have distinct radiographic features where radiomics solution can be used to delineate features and extract feature values, then machine learning methods can be used to build the COVID-19 recognition and diagnosis model.

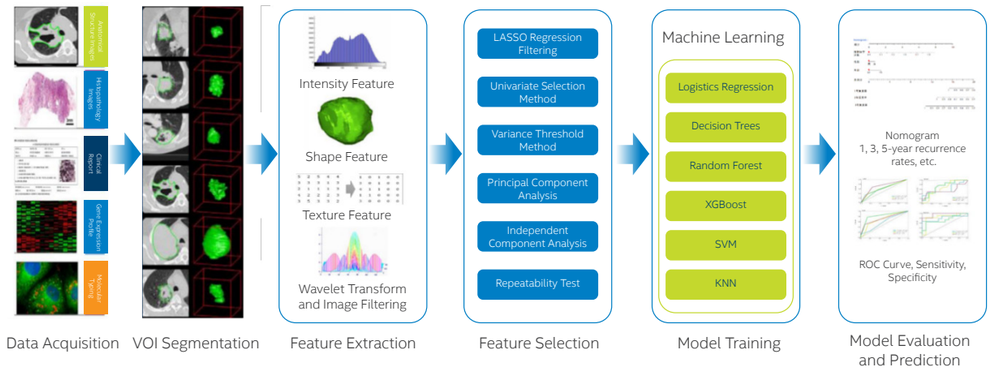


Figure 2-5-2 Basic Analysis Workflow of Radiomics

[50] Data cited from the **"Report on the Nutrition and Chronic Disease Status of Chinese Residents (2015)"** published by the Bureau of Disease Prevention and Control of the National Health and Family Planning Commission

[51] Radiomics was first mentioned by the Dutch scholar Philippe Lambin in his paper **Radiomics: Extracting more information from medical images using advanced feature analysis:** https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4533986/

More machine learning methods are being applied in the healthcare industry

■ Decision Tree and Random Forest Algorithms

A decision tree is a tree-structured supervised learning model that will test each feature, produce different results and branch, and each branch then will test relevant features and continue branching until the branch does not satisfy the splitting criteria, thereby inferring the final classification results.

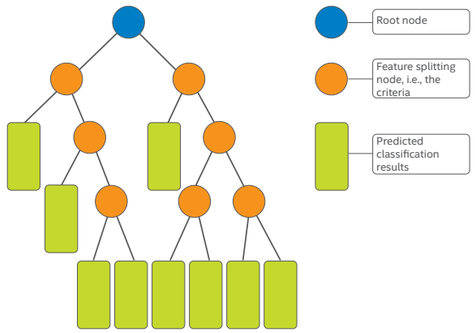


Figure 2-5-3 Decision Tree Model

In the prediction of heart disease or epidemic infection, for example, the model selects features such as age, blood pressure, medical history and others as the criteria. By using a certain amount of sample data for training, we can achieve the best feature splitting node through training, and calculate feature values corresponding to the disease, thereby obtaining the prediction model. The patient's feature values are then input into the trained model to infer the probability of disease.

The depth-increasing of the decision tree is prone to overfitting, and the random forest algorithm can effectively solve this problem. In simple terms, random forest, as an integrated learning method, constructs a forest of decision trees in a random manner, and when a new sample is included in the random forest, each decision tree will make a judgment, calculate the classification of the sample, and then "vote" to get the classification of the predicted sample. Random forest can be used to solve both classification and regression problems, which happen to be some prioritized qualitative and quantitative problems in clinical diagnosis, such as population-specific screening.

■ Logistics Regression Algorithm

The Logistics Regression (LR) algorithm is one of the common machine learning algorithms, which introduces the **Sigmoid** function on the basis of linear regression and maps the linear regression **(-∞,+∞)** value domain to **(0-1)**, thus predicting the probability of a certain disease. For example, in the infectious disease prevention and control model, with viral infection set to **a=1** and uninfected set to **a=0**, the feature value **b** (age, gender, medical history, travel history, exposure history, etc.) from **N** independent samples are introduced into the following target function:

$$G(a=1|b;\theta)=\frac{1}{1+e^{-\theta^T b}}$$

Maximum likelihood is then used to solve for the maximum value and regular term optimization is introduced to penalize large parameters to avoid overfitting, thus calculating the optimal parameter values. Ultimately, by training with sample data, we can obtain a probability model for determining whether an infection is present or not.

The LR model allows discretization of continuous numerical features and easy model iteration, and is highly robust. It is more helpful to improve the expression of the model after performing feature crossover for discrete vectors.

■ Several Boosting Algorithms

Integrated learning is a type of machine learning that can generate a strong classifier by combining a series of weak classifiers. When there is a strong dependency between the weak classifiers, as shown in Figure 2-5-4, and each weak classifier has a serial relationship with each other, the integrated learning is also called Boosting algorithm. Typical Boosting algorithms include Adaptive Boosting (AdaBoost) and Gradient Boosting Decision Tree (GBDT) algorithms.

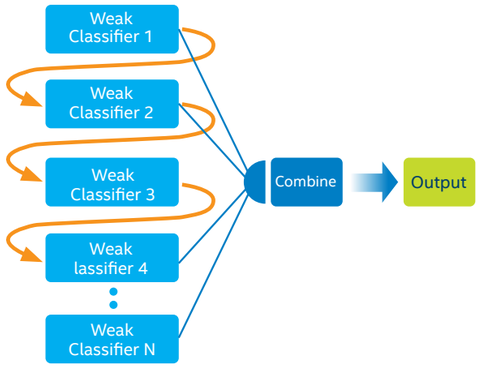


Figure 2-5-4 Integrated Learning Method in Machine Learning

The AdaBoost algorithm adopts the idea of iteration and typically uses a single-level decision tree as a weak classifier. Only one weak classifier will be trained per iteration, and then the trained weak classifier will participate in the next iteration. After **N** iterations, there will be **N-1** trained weak classifiers with constant parameters and an Nth iterator that needs to be trained, and the final performance of the model will depend on the combined performance of the **N** weak classifiers. During the training process of the AdaBoost algorithm, each iteration changes the sample weights and the corresponding weak classifier weights, so it can be constantly adjusted according to the characteristics of different weak classifiers.

The GBDT algorithm is a strong classifier assembled from a series of Classification and Regression Trees (CART). CART regression tree will continuously be split along the binary tree according to features. For example, if current tree node **J** will be split based on **a**, the number of features, then samples with the number of features less than **b** are split into left subtree, while samples with the number of features greater than **y** are split into right subtree:

$$J_1(a,b)=\{X|X^{(a)}\le b\}\ (\text{left subtree})$$

$$J_2(a,b)=\{X|X^{(a)}> b\}\ (\text{right subtree})$$

In essence, CART regression tree is to divide the sample space by the feature dimension. The objective function generated by typical CART regression tree is:

$$\sum_{x_i\in J_m}(y_i-f(x_i))^2$$

Like the AdaBoost algorithm, the GBDT algorithm also adopts an iteration approach and its target function can be expressed as:

$$F(x)=\sum_{i=0}^{M}a_i f_i(x)$$

However, compared with the AdaBoost algorithm, the GBDT algorithm will generate a residual between the predicted value and the actual value at each round, which will be used for the prediction at the next round. Finally, all the predicted values are added together to obtain the prediction result. The GBDT algorithm is highly robust, which is especially important for complex data acquisition in epidemic prevention and control scenarios.

XGBoost, which has received much attention in recent years, is an excellent extension and efficient implementation of the GBDT algorithm. Its core idea is to generate new split trees through continuous feature splitting. Each tree added is actually learning a new function to fit the residual of the last forecasting. Therefore, the objective function of XGBoost can be defined as:

$$y_i=\sum_j q_j x_{ij}$$

When there are **k** samples, the forecasting result of model at the **N-th** round is:

$$y_i^{(N)}=\sum_{k=1}^{N}f_k(x_i)=y_i^{(N-1)}+f_N(x_i)$$

Compared with AdaBoost and GBDT algorithms, XGBoost has the following benefits:

- XGBoost supports parallel computing and can make full use of the multi-thread capability of processor. Especially when running on Intel® platform, it can effectively utilize the powerful vector computing power brought by the latest instruction set such as Intel® AXV-512;
- XGBoost introduces a regularization item in its cost function, which can effectively control the complexity of the model and prevent the model from over-fitting;
- XGBoost supports column subsampling, which can not only prevent over-fitting, but also reduce computational complexity.

■ LASSO Algorithm

It is well known that models are prone to overfitting when there are many sample features and a relatively small number of samples. Typically, the overfitting problem can be mitigated in two ways. The first is to reduce the number of features, and the second is to reduce the order of magnitude of the feature parameter **w** by using regularization. Regularization refers to the selection of a model with a small average loss function and a low model complexity. Thus, the LASSO algorithm aims to optimize the introduced regularization term (a monotonically increasing function that represents the model complexity), therefore, the larger the regularization term, the lower the model complexity and the lower the probability of overfitting.

The regularization term can be a norm of model parameter vector, and typical norms include L1 norm and L2 norm. The LASSO algorithm is the regularization of the L1 norm and its optimization goal can be expressed as:

$$\frac{min}{w}\sum_{i=1}^{m}(y_i-w^T x_i)^2+\lambda||w||_1$$

where the regularization parameter **λ > 0**. Also, the L1 norm regularization is easier to obtain a sparse solution, i.e., its solution to w will have fewer non-zero components. The LASSO algorithm can be solved using the Proximal Gradient Descent (PGD) method[52].

LASSOCV, on the other hand, is an iteratively fitted LASSO linear model along the regularization path, which is based on the LASSO method with K-Fold cross validation to automatically find the optimal model. Cross validation is a common approach used in machine learning to build models and verify model parameters. When there is insufficient data, LASSOCV will slice the given data and then combines the sliced datasets into training and test sets, on which training, testing, and model selection are repeatedly performed. The K-Fold cross validation is a cross validation method that slices the dataset into **K** subsets, trains **K** times (taking one subset as the test set each time), and obtains the tuner group with the lowest cross-validation error.

The LASSO algorithm is now widely used in compressed sensing, image processing, and trend analysis.

* For more details on the LASSO algorithm, see Section 11.4 of the **Machine Learning** (Watermelon Book) by Professor Zhou Zhihua.

[52] The above LASSO-related algorithms are described in part with reference to Section 11.4 of the **Machine Learning** (Watermelon Book) by Professor Zhou Zhihua.

■ PCA Algorithm

The PCA algorithm is a widely used algorithm for data dimensionality reduction. Dimensionality reduction is the pre-processing of high-dimensional feature data to speed up data processing and improve machine learning model efficiency by removing noise and minor features from the high-dimensional data. In simple terms, as shown in Figure 2-5-5, the PCA algorithm works by rotating the baseline (red line) on the data coordinate axis (blue coordinate axis) all the way to the direction where the data variance is the largest (triangular data projection on the baseline is the largest), and then using feature value analysis to determine the number of principal components that need to be retained, thus achieving data dimensionality reduction.

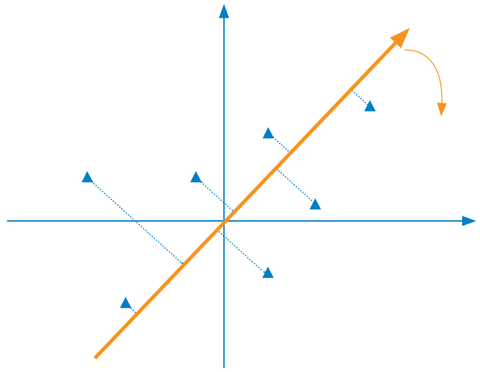


Figure 2-5-5 PCA Algorithm Mapping Diagram

In general, assuming there are **X** rows of **Y**-dimensional raw data, the basic workflow of the PCA algorithm is as follows:

- Read in the data matrix, and arrange the data by column to generate a matrix **Z** with **Y** rows and **X** columns;
- Perform zero-mean on each row of **Z** (representing an attribute field), i.e., subtracting the mean value of that row;
- Calculate the covariance matrix;
- Calculate the eigenvalues of the covariance matrix and the corresponding eigenvectors;
- Arrange eigenvectors in rows from top to bottom according to the size of the corresponding eigenvalue;
- Retain the first **W** rows to generate a matrix **D**;
- Transform the data into a new space constructed by eigenvectors, where **Y=DX** is the data after dimensionality reducing to k dimensions.

The PCA algorithm has the following advantages over other dimensionality reduction algorithms:

- Unsupervised learning, and not limited by parameters;
- Orthogonality between principal components, minimizing the interaction between data components;
- Low computing overhead and easy to implement;
- Effective in removing noise;
- Allowing data compression with minimal information loss.

Thanks to these advantages, PCA algorithm is now widely used in the exploration and visualization of high-dimensional datasets, as well as in data compression, medical/financial data preprocessing, and speech analysis.

■ Model Stacking Algorithm

When using machine learning, the generalization ability of a single model is often relatively weak, but model stacking can be used to combine the advantages of multiple models to improve prediction accuracy. Common model stacking methods include weighted stacking, Stacking/Blending, and so on. The stacking is a hierarchical model integration strategy, in which each layer can use multiple base learners to build a separate model in that layer, and then the output of the previous layer is used as the input of the next layer for training, thereby obtaining the final prediction result by accumulating prediction in each layer. The solution achieves higher prediction accuracy by combining the advantages of different models.

Intel® Architecture Improves Efficiency of Machine Learning Methods

High-dimensional Machine Learning Models in Medical Applications

According to the 80/20 Rule, the traditional expert rule system can cover 80% of the population with its human experience and expertise. However, the 80/20 Rule, also known as the Vital Few Rule, means that in any set of affairs, the most important accounts for only 20% of them, while the remaining 80%, although a majority, is of secondary importance. However, the remaining 20% of the population would require several orders of magnitude more dimensionality if covered by the rule. In this case, if the medical data is mined by a machine learning model to increase the number of feature dimensions to millions or even billions, the 20% of long-tail users can be effectively covered.

In specific medical application scenarios, traditional expert rule systems may use medical test results alone to determine whether a user is diagnosed with diseases, or screen for diseases based on typical symptoms, such as fever or rash. However, when using machine learning models, more topological relationships, such as the user's own health record, and whether he or she is a high-risk group for a certain disease, can be used to form a high-dimensional combined feature, which can greatly improve the coverage and accuracy of disease determination and identification over the rule-based model, ensuring an improved recall while maintaining a high level of accuracy. Taking the epidemic prevention and control as an example, as shown in Figure 2-5-6, companies such as 4Paradigm have introduced discrete high-dimensional models that can increase the number of dimensions to tens of millions, or even hundreds of millions.

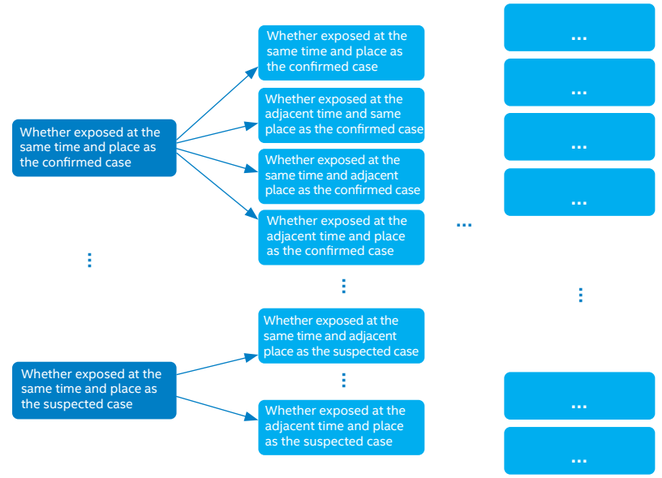


Figure 2-5-6 Feature-derived Engineering in Epidemic Prevention and Control

Building high-dimensional machine learning models can bring the following advantages:

- **High-dimensional features (rules):** each feature corresponds to a business rule, while the business rules are summarized by human and small in number (usually less than a thousand), so its ability to express the real world is poor; however, the number of rules (features) used in high-dimensional models reaches millions, which is much larger than that in the general business model, and can greatly improve the accuracy of prediction and recognition;
- **High-dimensional models (non-linear):** linear models, including rule models, are weak in expression, and non-linearization of linear models needs to use methods such as kernel function, manual discretization, and feature combination, which requires a lot of manual work before learning. In contrast, the tree model can produce highly nonlinear models with input from massive amounts of real data. The dimension is exponentially related to sample size and is more expressive of the real world.
- **High-dimensional model stacking:** Although high-dimensional models are highly expressive, unrestricted increase of dimensions may lead to overfitting. When using model stacking, each classifier will capture different aspects of high-dimensional data, thereby expressing higher dimensions. In addition, the implied regularization of model stacking can prevent overfitting.

Intel Provides More Powerful Hardware Infrastructure Support for High-Dimensional Machine Learning Models

High-dimensional machine learning models, unlike general algorithmic models, often build a giant pyramidal data matrix whose underlying data dimensions may be in the hundreds of millions, so while they bring advantages, they have an urgent need for general-purpose computing power and massive amounts of memory, as well as higher performance requirements for the underlying hardware infrastructure.

The higher clock frequency, more processor cores, and more threads of Intel® processors provide greater computing power to support high-dimensional models. The second-generation Intel® Xeon® Scalable processors feature up to 56 processor cores, 112 threads and a fully upgraded and optimized micro-architecture, as well as faster and more efficient caching for improved processing performance, and support for up to 36TB of system-level memory. Its integrated Intel® AVX-512 provides a broader range of vector computing capabilities to support the efficient execution of multiple algorithms of machine learning.

At the same time, the high-dimensional model means that the system must cope with massive amounts of data processing. Typically, when the dimension of data is in the millions, the file size reaches GB, while in the billions, the file size can be as large as TB. Machine learning systems produce a large amount of intermediate result data for model iteration regardless of the algorithm used for model computation and update, and the storage performance of these intermediate results will obviously limit the further improvement of training speed. Meanwhile, the epidemic simulation can't be delayed, and requires that the intermediate data shall not be lost in the event of an accident.

In traditional infrastructure facilities, DRAM memory is typically used to meet high performance storage requirements. But as the dimensions of the data reach a certain magnitude, more affordable and reliable hardware facilities are needed to provide storage capacity. Built on a unique 3D XPoint™ storage medium, Intel® Optane™ persistent memory is the obvious choice for both high performance and high capacity, and provides two important features that are required by high-dimensional machine learning models: high density and persistence. The former means a maximum memory density of up to 512 GB per DIMM slot, several times that of today's DRAM memory, while the latter allows the server to retain data even in the event of a power outage or reboot.

Use Cases

4Paradigm Supports Precise Epidemic Prevention and Control with High-dimensional Machine Learning Models

■ Background

The outbreak of COVID-19 not only poses a huge threat to people's lives, but also deals a blow to socio-economic development. In order to effectively fight the epidemic, in addition to isolating and treating patients and their contacts, it is more important to track the epidemic transmission path, screen high-risk groups, and then take effective prevention and control measures to avoid widespread outbreak. Meanwhile, it is also necessary to simulate and predict the development trend of the epidemic, and develop the prevention and control measures at the macro level to achieve precise epidemic prevention and control. These approaches rely on the acquisition and aggregation of a full range of epidemic-related feature data, as well as efficient processing and analyzing with big data and AI.

Data from the front lines of the epidemic prevention and control indicate that, the COVID-19 is characterized by a high transmission rate and a long incubation period, and the person with COVID-19 may transmit the virus to any person who he or she has encountered at any place. Due to the discrete distribution of location, people, time, etc., such feature data is often high dimensional sparse data. Due to the inability to effectively express the high-dimensional sparse data features, traditional machine learning algorithms is not good at processing such feature data, and are prone to overfitting which leads to reduced accuracy. Meanwhile, millions and even billions of data dimensions place higher demands on the computing power and storage performance of the system.

To address this challenge, 4Paradigm, by leveraging its extensive experience in machine learning, builds a new data-driven digital twin system for COVID-19 with high-dimensional machine learning algorithms on advanced Intel® hardware products, which can provide accurate predictions for virus transmission path analysis, precise prevention and control, and epidemic simulation and decision making. The solution has proven its effectiveness in facilitating departments at all levels to carry out epidemic prevention and control faster and better.

■ Solution and Results

If the fight against COVID-19 is a war, 4Paradigm, Intel, and other partners working together to develop highly accurate predictive solutions undoubtedly engage in a "Defensive Battle" with AI and big data as their weapons. From the analysis of virus transmission in the early phase of prevention and control, to the accurate screening of susceptible populations, to the later simulation of epidemic, and from predicting the rate of infection in a region, to rehearsing the overall prevention and control decision for a province or city in advance, to supporting decision makers in making policy and influencing the development of the epidemic, each and every action must be "fast and accurate".

According to the different phases of epidemic development and the needs of the actual scenario, the system team developed three sets of applications, corresponding to tracking the transmission path, screening of high-risk populations, and simulating epidemic development.

■ Step 1 of precision prevention and control: tracking the transmission path

In the aftermath of an outbreak, analysis of virus transmission path is extremely critical, and the new solution provides the ability to accurately and efficiently analyze the path of virus transmission. This solution is different from traditional analysis methods that require manual analysis of patient information to find correlations, followed by simulation and field validation to find the source. As shown in Figure 2-5-7, the new solution establishes a network of potential transmission, uses a spectral clustering algorithm to slice the transmission network, and combines patient information with a search algorithm to obtain potential transmission paths, which is more responsive to complex scenarios, new situations and new information than traditional methods.

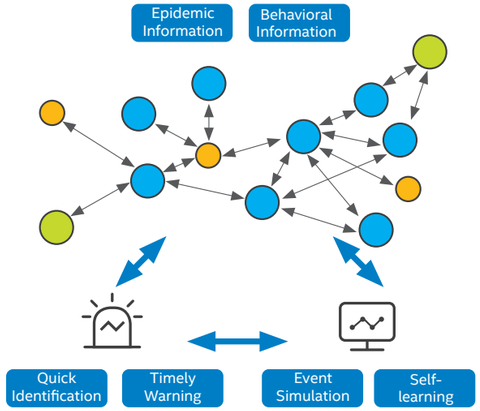


Figure 2-5-7 Transmission Path Analysis Solution

In addition, the new solution includes a self-learning event replay simulator that can learn from changes in the situation, making it possible to quickly identify potential infection path and facilitate epidemic prevention and control departments quickly block the spread of epidemic.

■ Step 2 of precision prevention and control: screening of high-risk populations

Blocking the spread of epidemic at the source is the first step in precise epidemic prevention and control, but effective screening measures are also required in order to specifically monitor or isolate potentially high-risk populations. Traditional screening method uses rule-based expert systems to build models, in which human experience and expertise are used to determine whether a user have been exposed in the same place and at the same time as the confirmed or suspected case, and to perform inference to make an analytical conclusion. However, when using this method, there are fewer data feature dimensions and fewer criteria for making determination, which may lead to a less accurate model. To improve accuracy, there must be enough feature data to support model training, because the feature dimensions summarized by human experience are far from sufficient.

To this end, 4Paradigm combines the original features and introduces temporal features, and derives high-dimensional sparse combined features from key information such as the topological relationship between locations, the activity level of the user, and whether the user is a high-risk user's co-traveler or co-traveler at different place and time, thus increasing the number of effective feature dimensions to millions or even billions. The new solution uses a total of tens of millions of data for modeling, in which there are 11 original features, 59 first-order features derived by feature engineering, and the feature combination is up to 5th order.

In addition to increasing the number of feature dimensions, 4Paradigm has carefully selected machine learning algorithms and the associated hardware infrastructure. Traditional machine learning algorithms have many shortcomings in dealing with high-dimensional sparse features. For example, although traditional decision tree algorithms can be used to handle large-scale feature data and well represent the real world, they are more suitable for handling dense data. Therefore, overfitting may occur and result in reduced accuracy if the data is too sparse. In contrast, the LR algorithm is good at handling sparse features, but when the number of features is too large, the training speed of model may be extremely slow.

To this end, the new solution adopts a model stacking approach, which means that, it uses a hierarchical model integration framework to build a separate model for each layer, and then uses the output of the previous layer as the input training set for the next layer, thus perfectly combining the advantages of gradient boosting decision trees and logistics regression algorithms and building a high-dimensional machine learning stacking algorithm specifically designed for high-dimensional sparse features to effectively improve model accuracy.

Meanwhile, the new solution introduces the second-generation Intel® Xeon® Scalable processors and Intel® Optane™ persistent memory to unleash the potential of algorithms. The second-generation Intel® Xeon® Scalable Processor provides high-frequency and multi-core computing capability for high-dimensional data features, enabling the trillion-dimensional model to achieve accurate predictions with millisecond response under million-level throughput, while the Intel® Optane™ persistent memory effectively improves the overall storage performance and computing efficiency by virtue of its non-volatile nature, read/write performance as well as access latency close to that of DRAM memory.

To verify the performance of the solution, 4Paradigm calculates the accuracy and recall of TOP K on a test set of 100,000 data samples and compares it to the theoretical upper limit. As shown in Table 1, the actual accuracy of the solution is very close to the theoretical upper limit, and the accuracy of the first 1,000 predictions even reaches the upper limit, i.e., the 1,000 persons most likely to be infected, as predicted by the solution, are all actually infected. As K increases, the model recall gradually grows to 90% at K = 10,000, meaning that simply screening those 10,000 person can find 90% of infections.

TOP K	Theoretical upper limit of accuracy	Actual accuracy	Theoretical upper limit of recall	Actual recall
1,000	1.0000	1.0000	0.3270	0.3270
2,000	1.0000	0.9036	0.6532	0.5908
5,000	0.6124	0.5102	1.0000	0.8334
10,000	0.3062	0.2757	1.0000	0.9004
20,000	0.1531	0.1432	1.0000	0.9357

Table 1 Test Results of 4Paradigm's Screening Model

■ Step 3 of precision prevention and control: simulating the epidemic development and assisting in decision-making and prediction at macro-level

Understanding the trends of epidemic and predicting the turning point are key for policy makers to effectively promote epidemic prevention and control. Undoubtedly, if the epidemic model is more accurate, then more targeted measures and manpower arrangements will be made by all levels of government, epidemic prevention and control institutions and medical organizations, and the allocation of medical resources will be more effective, which will also help more patients to recover as soon as possible, and make more reasonable arrangements for production resumption.

Internationally, classical infectious disease transmission models such as the SIR infectious disease model, SEIR infectious disease model, Gaussian process regression algorithm, and the SARS infectious spread model are used to predict and simulate epidemic. For example, Figure 2-5-8 shows the classical SEIR (Susceptible-Exposed-Infectives-Resistances) infectious disease model. It summarizes the four phases of a general infectious disease, that are Susceptible, Exposed, Infectives and Resistances, and uses the variation coefficients a, b, and c in these four phases to calculate a graph of epidemic development.

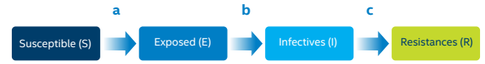


Figure 2-5-8 SEIR Model

The assumptions of the SEIR model are that the total number of individuals in a region remains constant (i.e., the mobility of people is not taken into account) and that no random factors (e.g., prevention and control measures) are included, but because the model lacks parameters such as human intervention factors, it cannot reflect the effect of prevention and control measures in real time, and the predicted value may be larger than the actual value, resulting in its inability to predict the turning point.

The way the population moves, the grid control measures, etc. will all have an impact on the model. For example, because of the way the air circulates, the infection rate in airplanes is much lower than that in trains, and in communities that pay more attention to protection against infectious diseases, the infection rate is lower, and so on. Traditional algorithms rarely take into account macro factors such as population movements and control measures, so the trend they predicted or simulated are not accurate enough to effectively support the decision-making related to prevention and control in the actual fight against epidemic.

To this end, 4Paradigm uses high-dimensional machine learning techniques and multi-dimensional data to develop a new solution. As shown in Table 2, the new solution is built on multiple data features such as the number of confirmed, suspected, cured cases, and deaths in the country, the number of population in each province and municipality, the number of migration between provinces and municipalities, and the population activity level within each province and municipality, and uses high-dimensional machine learning methods to build a more fine-grained and realistic digital twin system at the provincial, city, district and county levels. The system model fully considers the impact of various unexpected factors in a complex environment, such as traffic control, drug delivery time, and regional control, on the development of the epidemic, and can accurately predict and simulate the development trend of the epidemic.

	Features	Volume of data (continuously increasing)
Statistics of national epidemic (by province and municipality)	Confirmed cases	14K+
	Suspected cases	
	Cured cases	
	Deaths	
Population by province and municipality	Population today by province and municipality	14K+
Migration between provinces and municipalities	Population migrating today across provinces and municipalities	30K+
Population activity level within province and municipality	Population traveling today within province and municipality	30K+
	Travelled distance today within province and municipality	

Table 2 Partial Features of the Epidemic Simulation Model

The new epidemic simulation model built on high-dimensional data is shown in Figure 2-5-9, in which the left and right sides show the model as of March 25, 2020. The blue curve represents the cumulative actual confirmed cases, while the red curve represents the cumulative confirmed cases predicted by simulator. The blue bar is the number of isolated cases of the day in the infected organization, the orange bar is the number of confirmed cases of the day in the isolated organization, and the green bar is the number of new confirmed cases of the day.

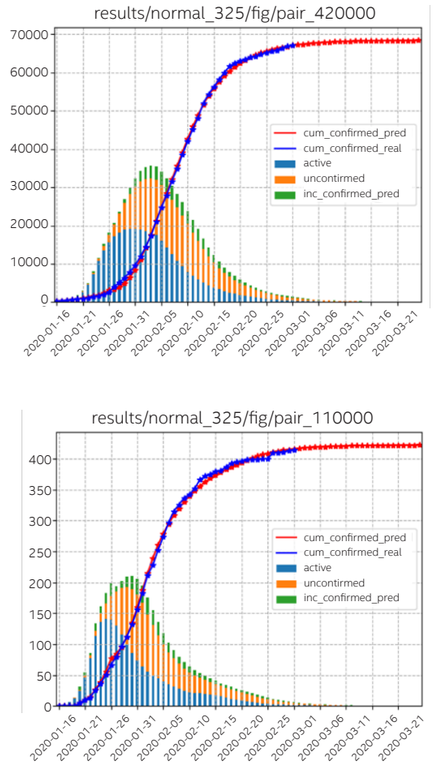


Figure 2-5-9 Epidemic Simulation Model Built on High-dimensional Data

It can be seen that, the confirmed cases simulated and predicated by the model are highly consistent with the actual confirmed cases. Also, the simulation and predication perform very well in terms of accuracy compared to the improved SEIR model. From the provincial data, the Root Mean Squared Error (RMSE) of the new solution is reduced by more than 90%[53] compared with the improved SEIR model, and enables a better fit.

By using a more accurate epidemic simulation model, as shown in Figure 2-5-10, epidemic prevention and control departments at all levels can use the system to make simulation and prediction before key decisions are made, thus providing a reliable and strong data support for the formulation of prevention and control policies.

Figure 2-5-10 4Paradigm's Epidemic Simulation System

4Paradigm Builds a Closed-Loop Management Solution for Chronic Disease Prevention and Management

■ Background

Chronic diseases have taken a huge toll on people's quality of life and our society and economics, and the most effective measure to combat them is effective prevention. As shown in Figure 2-5-11, chronic disease prevention and management can be summarized as the following four steps:
(1) Establishment of health records for high-risk groups such as the obese, smokers, middle-aged and elderly, and those with disease history;
(2) Risk assessment for chronic diseases such as diabetes, cardiovascular diseases, strokes and hypertension in a scientific way;
3) Personalized health intervention programs, such as exercise programs, diet programs, and so forth;
4) Long-term tracking of intervention effects to determine risk trends and adjust intervention programs.

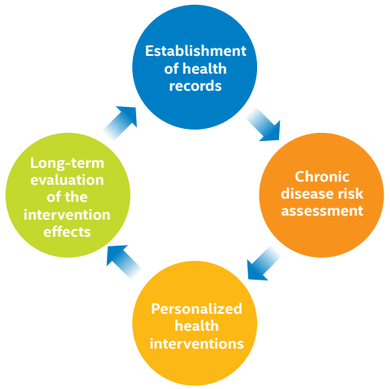


Figure 2-5-11 The Four-Step Approach to Chronic Disease Prevention and Management

In the past, all of the above work required the professional advice of experienced medical professionals, health experts, nutritionists and exercise specialists. But at a time when medical resources are increasingly strained, it is clearly not practical to provide universal specialist services to the public. In addition, even expert services rely on personal experience to make judgments, making it difficult to satisfy precise and individual needs. So, how to use technologies to provide more residents with high-quality chronic disease prevention and management services has become a new challenge for many medical and health institutions and high-tech enterprises.

Combining rich medical data with machine learning to implement risk assessment, personalized health intervention, and evaluation of intervention effect has become an effective way to address chronic disease challenges. Following the above concept, 4Paradigm cooperated closely with Ruijin Hospital, affiliated with Shanghai Jiao Tong University School of Medicine. By utilizing the expertise, clinical experience and the world's largest sample database of metabolic diseases of Ruijin Hospital, and using AI Sage, 4Paradigm's machine learning platform, they developed the Zhining series of chronic management products on Intel's advanced hardware and software products, including the Zhining Chronic Disease Management All-in-One Machine, Chronic Disease Management Cloud System, Ruining Zhitang, Ruining Zhixin, Chronic Disease Management Follow-Up Box, and Miniature Health Robot, which facilitate medical and health institutions implement prevention and management throughout the whole process of chronic disease.

■ Solution and Results

As shown in Figure 2-5-12, the chronic disease prevention and management services developed by 4Paradigm mainly consists of the Zhining Chronic Disease Management All-in-One Machine and the Chronic Disease Management Cloud System, which allow users to log into the Chronic Disease Management Cloud System, use the Chronic Disease Management All-in-One Machine to perform intelligent testing, and establish their own chronic disease management files. After the test data is uploaded to the cloud platform, a machine learning model is used to accurately assess the risk of chronic diseases and combine it with multi-disease risk factor analysis to provide a scientific and personalized health intervention program; meanwhile, the solution also makes use of the "Health Paradigm" WeChat public account and Mini Program to provide users with intelligent reminder and tracking management services, thereby achieving long-term evaluation and management of the intervention effects. Through the above closed-loop system, users can obtain a comprehensive chronic disease management service that integrates physical examination, screening, intervention and management.

Figure 2-5-12 Closed-loop of Chronic Disease Prevention and Management

Currently, Zhining Chronic Disease Management All-in-One Machine is able to perform nearly 20 different tests, including blood pressure, blood glucose, uric acid, cholesterol, blood oxygen, electrocardiogram, fat rate, metabolic index, muscle content (%), water content (%), body temperature, and so on. The cloud system adopts a semi-supervised multi-tasking SMT-GBDT machine learning algorithm, based on the world's largest and most up-to-date metabolic disease sample database, in order to build a highly accurate screening model for chronic diseases in China, including diabetes, cardiovascular disease, stroke, hypertension, and so on. In practice, the performance of the model is far better than current standards (including those of developed countries such as the United States and Finland and those of the Chinese Medical Association), i.e. its prediction accuracy is 2 to 3 times better than that of clinical gold standard currently used by medical professionals[54].

Behind the advanced algorithm is the ultra-high-dimensional machine learning approach developed by 4Paradigm. From the previous description of high-dimensional models, it is clear that the higher the dimensionality of the model in machine learning, the better the learning ability. In traditional chronic disease management system that relies on the experience of experts, only thousands of expert rules can be summarized from a doctor's many years of experience. As a result, it is increasingly apparent that, such system is not able to address today's diverse lifestyles and satisfy higher requirements for chronic disease management.

The cloud system responds to this challenge by slicing the test data and building ultra-high-dimensional machine learning capabilities. In the data pre-processing phase, the system first models data of whole samples. In the feature engineering phase, basic information such as test result data and user information are extracted, and combined with the subject's historical medical records, family medical history and other diversified features. The ultra-high-dimensional machine learning algorithm will perform ultra-high-dimensional combinations and derivations of the original data fields by using the powerful computing power of server clusters based on Intel® processors, thereby generating tens of millions or even hundreds of millions of features.

Compared with traditional machine learning algorithms, 4Paradigm's GBDT machine learning algorithm outperforms decision tree models in terms of model accuracy and the ability to use discrete features. As shown in Table 3, 4Paradigm's GBDT algorithm takes into account the requirements of model accuracy and the need to prevent model overfitting, and the number of modeling samples and the number of input features it supported are also significantly higher than traditional integrated learning decision tree algorithms.

	Traditional Decision Tree Algorithm	4Paradigm's GBDT Algorithm
Number of trees	Single tree	Multiple trees
Model accuracy	Too deep tree may lead to over-fitting, and it's hard to achieve accuracy and prevention of over-fitting at the same time	Iterate with many simple trees, which is not easy to over-fitting
Number of samples	Millions	Hundreds of millions
Input Features	Thousands	No limit, depending on the size of platform nodes
Ability to use discrete features	Inability to handle large scale discrete features	Enables the processing and use of large-scale discrete features

Table 3 Comparison of 4Paradigm's GBDT Algorithm with Traditional Decision Tree Algorithm

[53] Data tested and measured by 4Paradigm.

[54] Data citied from product manual of **4Paradigm Zhining Chronic Disease Management All-in-One Machine**.

To improve the effectiveness of chronic disease prevention and management, 4Paradigm has introduced Intel® products to improve efficiency throughout the closed-loop process. On the one hand, the introduction of the second-generation Intel® Xeon® Scalable processors gives the platform necessary computing power to address the challenges of processing trillions of high-dimensional data. The Intel® AVX-512 technology integrated into the processor also accelerates the model prediction process with its powerful vector computing ability. On the other hand, Intel® Optane™ SSDs combine high throughput, low latency, high quality of service and high durability to provide a high-quality data storage infrastructure for the platform.

The new solution has achieved good performance in a number of healthcare organizations. As shown in Figure 2-5-13, in a comparison of the predictive performance of several common chronic diseases, 4Paradigm machine learning algorithm's Area Under Curve (AUC)[55], one of the model evaluation indicators, is better than those of the control group.

Intel® Distribution for Python Helps Huiyi Huiying Improve the Efficiency of Radiomics Feature Selection

■ Background

Radiomics can be used to further mine the information contained in medical image data, and facilitate medical organizations find subtle lesions earlier and faster, thus eliminating cancer and other malignant diseases at an early stage, greatly reducing the suffering of patients, and effectively improving the usage efficiency of medical resources to enhance the health of the entire population. As an active pioneer of radiomics and its solutions in China, Huiyi Huiying is providing medical organizations with "full-cycle" and "one-click" imaging big data scientific analysis capabilities through its products and platforms such as AI All-in-One Machine, equipping medical organizations with necessary tools for the application of radiomics.

As can be seen from the previous section ("Radiomics Applications" on page 56), the basic process of radiomics is divided into steps such as data acquisition, VOI segmentation, feature extraction, feature selection, model training, and model evaluation and prediction. Since the idea of radiomics is to extract as many data features from medical images as possible, problems such as possible Curse of Dimensionality need to be properly addressed. The Curse of Dimensionality in machine learning means that for a given sample size, spatial data becomes more sparse as the number of input dimensions increases, which can seriously affect the predictive performance of the model. To solve this problem, it is necessary to select an appropriate algorithm to reduce dimensionality of data features during the feature selection phase.

When deploying radiomics solutions, healthcare organizations need to train large dataset to enable them making more accurate predictions about patient image data, which requires a feature selection step with higher processing efficiency. Therefore, it is critical to equip the solution with hardware infrastructure that has higher processing power and to make targeted tuning based on algorithm characteristics. To satisfy this demand, Huiyi Huiying not only introduces the 2nd generation Intel® Xeon® Scalable Processor as the powerful processing engine of the solution, but also adopts the Intel® Distribution for Python to improve the efficiency of feature selection algorithms such as LASSOCV and PCA.

■ Solution and Results

In radiomics-based medical image processing system, algorithms such as LASSOCV and PCA are the most commonly used algorithms for the feature selection step, which can effectively facilitate the system mitigate the common Curse of Dimensionality problem in radiomics processes, allow the system to minimize information loss while compressing the data, and facilitate data visualization for more intuitive information presentation.

Huiyi Huiying's AI All-in-One Machine is equipped with the 2nd generation Intel® Xeon® Scalable processors. The processor not only integrates more processor cores and threads and a fully upgraded and optimized micro-architecture, but also features more cache for improved processing performance and supports up to 36TB of system-level memory; the built-in Intel® AVX-512 provides powerful vector computing power for model training and prediction tasks in radiomics solution, enabling more efficient execution of LASSOCV and PCA algorithms.

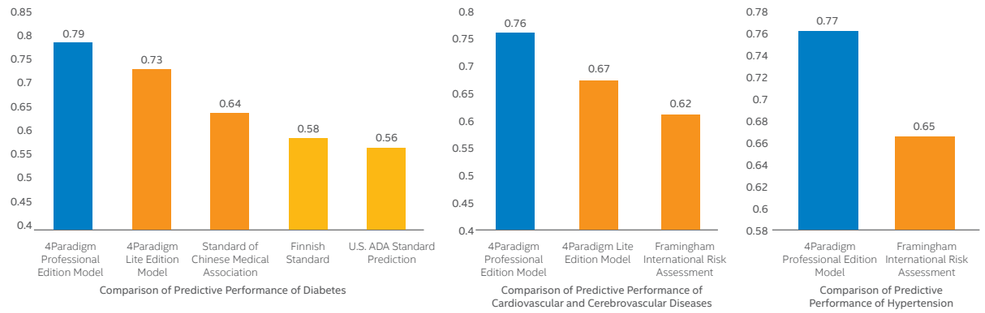


Figure 2-5-13 Comparison of Predictive Performance of Several Common Chronic Diseases[56]

[55] Please refer to Wikipedia for the definition of AUC values: https://en.wikipedia.org/wiki/Receiver_operating_characteristic#Area_under_the_curve
[56] Data citied from product manual of **4Paradigm Zhining Chronic Disease Management All-in-One Machine**.

Meanwhile, Huiyi Huiying also worked with Intel to optimize the algorithm execution language Python. The Intel® Optimization for Python, powered by Intel, incorporates more support for Intel® performance libraries such as Intel® MKL, and has the latest built-in vectorization instructions. More importantly, it also provides excellent support for the Scikit-learn (sklearn) library.

The Sklearn library is one of the most commonly used third-party libraries for machine learning, encapsulating common machine learning algorithms such as LASSOCV and PCA, as well as providing methods such as K-Fold cross validation for easy invocation by users. The environment configuration commands in Intel® Optimization for Python are as follows:

```
1. os.environ["KMP_BLOCKTIME"] = "0"
2. os.environ["USE_DAAL4PY_SKLEARN"] = "YES"
```

where **KMP_BLOCKTIME** is the time to wait before a thread finishes executing the current task and goes to sleep, here it is set to 0 milliseconds, and **USE_DAAL4PY_SKLEARN** is set to use the SKLEARN library.

Compared to native Python, Intel® Optimization for Python has tremendous efficiency gains in the actual execution of feature selection. As shown in the top graph of Figure 2-5-14, in the LASSOCV algorithm workload using K-Fold 10 cross validation with all radiomics features checked, Intel® Optimization for Python executes 2.12 times faster than native Python. As shown in the bottom graph, in the LASSOCV + PCA algorithm workload using K-Fold 10 cross validation with all radiomics features checked, Intel® Optimization for Python executes 2.08 times faster than native Python.

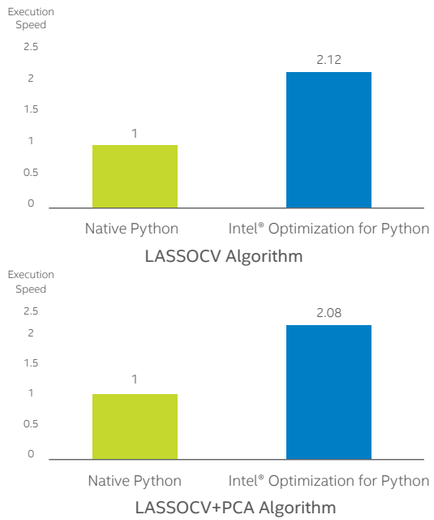


Figure 2-5-14 Performance Comparison between Intel® Optimization for Python and Native Python[57]

[57] The test configuration is as follows: Processor: 2S Intel® Xeon® Gold 6252 Processor @ 2.1GHz, 24 cores 48 threads; Memory: 192GB DRAM memory; Storage: Intel SSDSC2BB48; BIOS Version: SE5C620.86B.02.01.0009.092820190230; Operating System Version: 18.04.1 LTS (Kernel: 4.15.0-91-generi); Native Python Version: Python 2.7.17; Intel® Distribution for Python Version: Intel-Python2019U5; Workload: Medical imaging classification training provided by Huiyi Huiying

Conclusion

Using machine learning to build more efficient and precise epidemic prevention and control, chronic disease prevention and management, and radiomics models, on the one hand, can effectively facilitate governments at all levels, epidemic prevention and control institutions and healthcare organizations understand the development trend of epidemic, implement reasonable and effective control measures, optimize the use of medical resources, and speed up the R&D of vaccines and antiviral drugs, thus protecting people from the threat of viruses and ensuring the smooth and safe operation of society and the economy; on the other hand, it can also reduce the suffering of patients and improve the health of the whole population through more effective disease prevention and detection methods.

For precise epidemic prevention and control, 4Paradigm, together with partners such as Intel and LAMDA Institute of Nanjing University, has carefully selected machine learning algorithms and the associated hardware infrastructure based on the characteristics of the epidemic data and models. For example, the Intel® architecture processors and Intel® Optane™ persistent memory provide powerful computing power and storage for high-dimensional machine learning models, effectively improving the performance of epidemic prevention and control solutions.

In chronic disease prevention and management, 4Paradigm, Ruijin Hospital affiliated with Shanghai Jiao Tong University School of Medicine and Intel have launched a closed-loop chronic disease prevention and management system to address the characteristics of chronic diseases. A range of Intel® architecture-based hardware and software products provide the computing and storage performance that enables the system to perform well in a variety of applications, including diabetes, hypertension and cardiovascular disease prediction.

Regarding the medical image processing solution based on radiomics, Huiyi Huiying and Intel have jointly built an AI All-in-one Machine based on machine learning, which uses the Intel® Distribution for Python to optimize the new medical image detection capability and facilitate medical organizations to detect early malignant tumor lesions. Meanwhile, the technology has also been applied to CT imaging of patients with COVID-19, allowing healthcare organizations to more accurately and efficiently detect suspected cases and reduce false positives.

Explore the Federated Learning-based AI Approach in the Healthcare Industry

Break Down Data Barriers to Improve AI Application Performance in Healthcare

Improve training performance with multi-source data

From the aforementioned, it can be seen that by using deep learning, machine learning and other methods, AI can effectively improve the efficiency of medical image processing, auxiliary diagnosis, disease prediction and drug R&D, helping doctors to understand diseases more comprehensively and accurately, so that patients can recover as soon as possible. In addition to choosing the right algorithm and equipping with sufficient computing power, the improvement of AI efficiency relies on providing more data for training, inference and validation to improve model accuracy. Especially in the applications of image segmentation and pathology slice analysis, the deep learning models used require a large amount of sample data for training in order to achieve better generalization and prevent overfitting.

To unveil the large number of features hidden behind the data pixels, deep learning models commonly used for medical imaging typically adopt a multilayer network such as the convolutional neural network, which has many hidden layers between the input and output layers, and the number of hidden layers determines the depth of learning. Some learning methods used by the model, such as backpropagation, compare the error between the output and the training data and thus calculate the error in the output, and then the associated hidden layer adjusts its weights to reduce the error rate.

As a result, deep learning typically requires a large dataset of different instances from which the model can learn the desired features and generate output with probability vectors. The more complex the image processed, the greater the amount of data required for training. Research has shown that, as shown in Figure 2-6-1, in traditional machine learning methods, AI performance initially grows with the amount of training data and later flattens out, whereas the performance of deep learning methods consistently grows with the amount of training data[58]. Therefore, providing more datasets of different instances for AI applications in the healthcare industry, especially deep learning-based AI applications, can effectively improve their performance.

As with most industries, the problem of "data silos" in healthcare is serious, where data from different organizations, and even different departments, are often not interconnected. However, fully interconnected data sharing will bring up an inevitable problem of how to protect data privacy and security. As we all know, data such as health status is extremely important personal information, and healthcare organizations definitely can't accept the risk of data breach from improper use of these data.

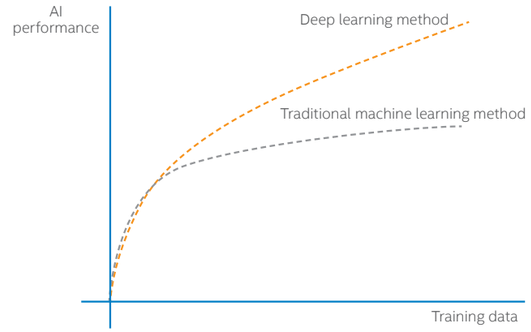


Figure 2-6-1 Effect of the Amount of Training Data on Different Learning Methods

In order to provide AI applications with safer and higher quality training datasets from more sources, many research and academic institutions have developed various collaborative learning methods, such as Institutional Incremental Learning (IIL), Cyclic Institutional Incremental Learning (CIIL), and the recently-renowned Federated Learning (FL).

When using the IIL method, participants involved in training will be arranged in sequence, and the training model is also passed in sequence, which means that the previous participant trains the model with its own data and then passes the results to the next participant who then re-trains the model with its own data. The CIIL is the iterated version of ILL. Both methods have some shortcomings in practice. Firstly, they both use a shared model, and the training model needs to be passed between different participants, which is likely to cause privacy leakage and data security problems; secondly, if the training data of each participant is too small, for example, each medical institution only provides the data of several patients, then the effect of collaborative learning cannot be effectively improved; finally, the above methods use a serial collaboration mode, which requires the model to be completely passed to the next participant, and has certain requirements on network performance.

In contrast to these two methods, the federated learning method uses a parallel, collaborative approach that allows all participants involved in collaborative learning to train the model locally using localized data, and then share the parameters of the trained model. The advantages of doing so are obvious; on the one hand, the training data and models of all participants remain local, which provides better protection in terms of data security and privacy protection; on the other hand, the parallel training mode can obtain cumulative training effects, effectively enhancing the training performance. In addition, when using the parallel collaborative method, the combination of data, model and training is similar to the distributed training, so the training efficiency is higher than that in the serial collaborative method.

[58] This view is summarized from Zhu, X. et al., Do we Need More Training Data?, https://arxiv.org/abs/1503.01508, March 2015; Shchutskaya, V., Latest Trends on Computer Vision Market, https://indatalabs.com/blog/data-science/trends-computer-vision-software-market?cli_action=1555888112.716; and Why go large with Data for Deep Learning?, https://towardsdatascience.com/why-go-large-with-data-for-deep-learning-12eee16f708?gi=ba92e606d0.

The key to build a federated learning system is the creation of a trusted data sharing approach for all participants. Currently, solutions based on Trusted Execution Environment (TEE) are becoming increasingly popular in the healthcare industry. Its core concept is to provide a secure, trusted and efficient computing environment on third-party hardware for different data sources. As shown in Figure 2-6-2, the training optimization results from different data sources A and B can be shared in a TEE environment created by the hardware on the right side, and the final optimization model can be generated. Among various TEE solutions, Intel® Software Guard Extensions (Intel® SGX) is more sophisticated and well-received by users.

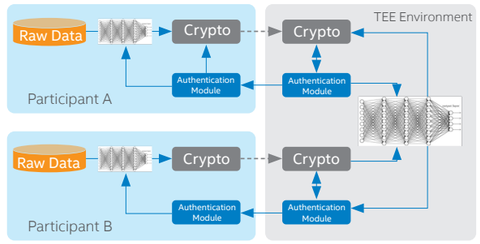


Figure 2-6-2 TEE Environment in Federated Learning

Federated Learning-based AI Methods

The federated learning can be divided into Horizontal Federated Learning, Vertical Federated Learning, and Federated Transfer Learning depending on the usage scenario, in which the Horizontal Federated Learning applies to scenarios where there are a lot of overlap on features but only a few on users in a dataset. It can slice the dataset by user dimension and retrieve data with the same feature but different user for training. For example, in the same pathology image processing, user data from different healthcare organizations can be trained in a horizontal federated learning manner.

The Vertical Federated Learning, on the other hand, applies to scenarios where there are a lot of overlap on users but only a few on features in different datasets. This method allows you to slice the dataset by feature dimension and retrieve data with the same user and different feature for training. In typical scenarios such as structured pathological diagnosis, data from the same group of users on different tests can be trained in a vertical federated learning manner. In scenarios where there are a few overlap on both users and features, the Federated Transfer Learning can be used to perform collaborative data training in a transfer learning manner, instead of slicing the data.

Taking the AI-powered pathology image segmentation as an example, healthcare organizations A and B each have a large amount of patient pathology image data, and for security and privacy concerns, these image data is stored in their respective data centers with high level isolation implemented by firewalls, and any direct access to the data will be denied.

In the process of training both datasets with federated learning, encrypted training is required with the help of a cooperative party C in order to ensure data confidentiality during the training process, as shown in Figure 2-6-3. The encryption training process is divided into the following steps:

1. The cooperative party C distributes the public key to A and B to encrypt the data to be exchanged in the training process;
2. A and B interact with each other to calculate the intermediate result of gradient in encrypted form;
3. A and B perform calculations with the encrypted gradient values, respectively, and provide the results to the cooperative party C. The cooperative party C calculates and decrypts the total gradient value by aggregating the results;
4. The cooperative party C sends the decrypted gradient back to A and B, respectively, and A and B update the parameters of their respective models with the decrypted gradient.

The above training steps will iterate until the loss function converges, the training process is complete and the final model is obtained. The parameters passed by the federated learning include:

- Typical hyper-parameters of deep learning architecture, such as Batch Size, Optimizer, Learning Rate;
- Epochs (EpR) in each round of learning. More Epochs can accelerate convergence, but the gain will diminish;
- Number of participants in each round of learning;
- The compression/pruning method used for model update.

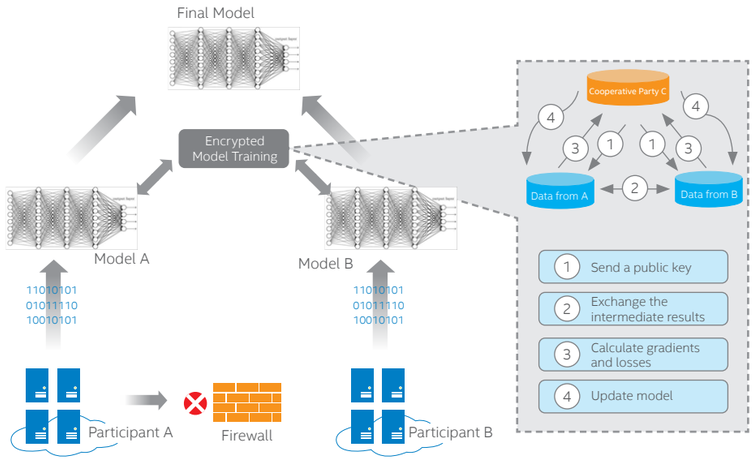


Figure 2-6-3 Basic Architecture of Federated Learning

Compared to a general distributed machine learning/deep learning approach, the federated learning approach has the following characteristics:

- Data is localized: participants train the global model using data they own;
- Every participant is involved in the learning process and model losses are controllable;
- The training process strikes a balance between privacy and security, and participants are able to co-construct the model without disclosing the underlying data and its encrypted form.

In addition, the federated learning also features a effect-based incentive mechanism, i.e., after the model is built through federated learning, the effectiveness of the model can be evaluated and documented by a permanent data logging mechanism. Participants who provide more high-quality data get better model results, and model results depend on the contributions that data providers make to themselves and others. The effects of these models are delivered to data sources in the federated effect-based incentive mechanism, so as to obtain federated incentives and continue to encourage other data sources to participate in the federated learning.

Based on the above characteristics, the federated learning can provide a cross-agency and cross-department data sharing and model training method for AI applications in the healthcare industry, helping to keep the private data from each data source localized, and build a learning model optimization mechanism without violating data privacy regulations by exchanging parameters in an encrypted way.

For source code of the federated learning, please refer to:
https://www.tensorflow.org/federated/

Intel® Software Guard Extensions (Intel® SGX)

Technical Briefs

As a typical implementation of TEE solution, Intel® SGX constructs a trusted Enclave in hardware (e.g., memory) through a new set of instruction set extension and access control mechanism that confines the security boundaries of data and applications to the Enclave itself and within the processor, enabling isolated operation between different applications. In addition, its operation doesn't rely on other hardware and software. This means that data security is independent of the software, operating system or hardware configuration, and even if the hardware driver, virtual machine or operating system is compromised by an attack, data breach and tampering can be eliminated, thus enhancing the security of application code and data.

Traditionally, protection for data privacy and security is mostly implemented at the operating system or software level, but when the operating system or software is "infected", the security of the data is in jeopardy. As shown in Figure 2-6-4, although applications can be protected against attacks from external hackers or malicious applications through security scans, firewalls, etc., but malware and malicious code that exploit operating system vulnerabilities can bypass these protections and directly attack critical private data.

Therefore, Intel® SGX provides users with enhanced security with the following key features:

- **Enhanced confidentiality and integrity:** the enclave works in an isolated hardware environment (Intel® architecture processors and memory with SGX support) and authenticates applications and data with keys, making it impossible to attack data even if privileged malware or malicious code is present in the operating system, BIOS, or virtual machine, and so on;
- **Smaller security attack surface:** Intel® SGX confines applications and sensitive data in protected hardware enclaves, which eliminates traditional attacks that malicious programs can launch from hardware, virtual machines and operating systems, resulting in a smaller attack surface for greater security;

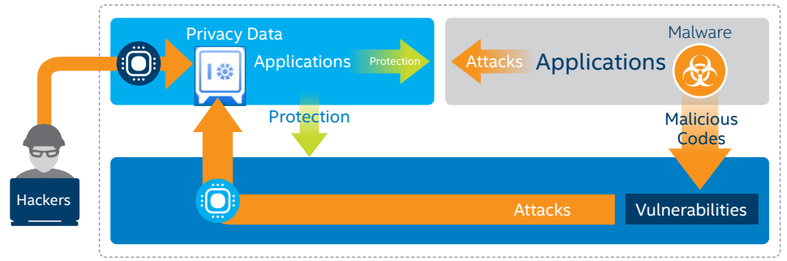


Figure 2-6-4 Internal Attacks on Applications

- **Remote authentication and control capabilities:** users can more securely provide keys, credentials and other sensitive data to enclaves by performing remote authentication.
- **Enhanced federated learning efficiency:** during federated learning performed on Intel® SGX, AI models and training data are deployed in protected hardware enclaves, which dramatically reduces communication and computing costs associated with application and data encryption and decryption, making learning more efficient;
- **Lower learning curve:** applications with Intel® SGX support can be developed, integrated and executed on a specific Intel® architecture-based processor platform, requiring only the installation of drivers and SDK adaptation, with no additional hardware or software environment, no changes to programming, and a lower learning curve;
- **More efficient implementation:** the TEE hardware solutions based on Intel® SGX run more efficiently than federated learning implementations that adopt techniques such as secure multi-party computation, homomorphic encryption, and so on.

Intel® SGX Installation and Configuration

Users can create an Intel® SGX-based solution by introducing the Intel® SGX SDK which provides:

- API
- Function library
- Document
- Sample source code
- tools

The latest Intel® SGX SDK can be obtained by clicking the following link:

Download link of SDK for Windows system

https://registrationcenter.intel.com/en/forms/?productid=2614

Download link of SDK for Linux system

https://01.org/intel-software-guard-extensions/downloads

Typical solution based on Intel® SGX

With Intel® SGX, healthcare organizations can build a variety of solutions based on their needs. The following section briefly describes a typical AI model training solution for multi-source data, which is built on Intel® SGX and has a cooperative party using central aggregator.

The solution architecture, shown in Figure 2-6-5, adopts a network that is composed of an aggregator enclave at the center and edge enclave deployed at each participant. The aggregator enclave and the edge enclave in each participant are the trusted regions that require privileged access and are constructed in memory by processor instructions provided by Intel® SGX.

In the solution, various parameters of the AI model are transmitted through the encrypted channel, while the training data, the plaintext AI model, and the AI algorithm are localized on each node. During the initialization process, each enclave first generates a public-private key pair, and the public key is registered with the aggregator, while the private key is stored in the respective edge enclave. When training begins, the aggregator first establishes a connection with the target enclave by using a symmetric encryption key. After the connection is established, the aggregator will first push the shared parameters of the model that need to be trained to each enclave in the encrypted form, and then each enclave will decrypt the model parameters and send them to the local AI training environment to train the local data. At the end of the training, the local AI training environment returns the shared parameters that are already trained to the local enclave.

The above transfer process between enclaves can be iterated in multiple rounds until satisfactory training results are obtained, and the solution can also evaluate the contribution of each participant to the training.

Since the above processes are implemented in enclaves, i.e., during the whole iterative cycle of the solution, the AI model parameters are passed through the encrypted channel and interacted in the encrypted enclaves, without any contact with external software and hardware, thereby establishing a secure and reliable internal loop. Meanwhile, AI models and training data are retained in various protected hardware enclaves, and only intermediate parameters need to be passed through the encrypted channel, which greatly increases the execution efficiency of federated learning. Intel® processors, especially the second-generation Intel® Xeon® Scalable processors, provide powerful computing performance to build enclaves, establish encrypted channels, and interact and aggregate intermediate parameters.

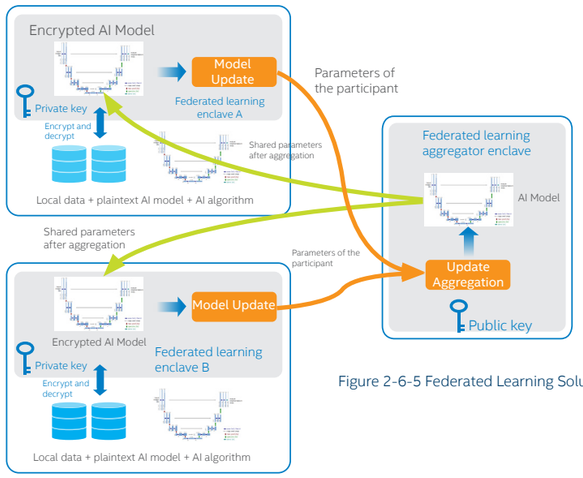


Figure 2-6-5 Federated Learning Solution with Intel® SGX

Research on the Application of Federated Learning in Medical Imaging

Background

Glioma is a common primary intracranial tumor located in the central nervous system, which can cause headaches, nausea and vomiting, epilepsy, and blurred vision, and is difficult to cure. Because gliomas tend to have different morphologies and discrete distributions, their localization and diagnosis is still difficult even with radiological imaging, histology and other medical detection methods. In radiographic imaging, for example, doctors are often required to distinguish precisely the biological properties of different tumor sub-regions in different tissue structures. The key to tumor detection is accurate segmentation of these sub-regions, which is why healthcare organizations now often use computer-assisted methods to process these medical images in order to determine their pathological characteristics and to evaluate subsequent treatment plans and prognosis.

Among these, Brain Tumor Segmentation (BraTS) is a commonly used computer-assisted segmentation method. The deep learning-based BraTS has been a hot topic in the field of medical image processing, and new approaches have been emerging in recent years at the International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI)[59].

There is a consensus that getting more data involved in training, such as creating publicly available, high-quality, multi-source datasets for benchmarking and quantitative evaluation, can further improve BraTS performance, but doing so in practice still faces significant challenges. On the one hand, sharing data to a centralized location still requires addressing system architecture and transmission efficiency issues; on the other hand, the availability of medical data is more limited than that of ordinary photographic images due to legal, privacy, technical and data ownership constraints.

Since 2018, Intel has been working with the Center for Biomedical Image Computing and Analytics (CBICA) at the University of Pennsylvania to jointly explore the application of federated learning in medical image processing, and achieved an effective implementation on BraTS, the results of which can be found in the paper **Ulti-Institutional Deep Learning Modeling Without Sharing Patient Data: A Feasibility Study on Brain Tumor Segmentation**. The following section provides a brief description on how the implementation uses the BraTS dataset and the federated learning method to build an effective image segmentation model by iteratively aggregating locally trained models on an aggregator, without sharing any patient data, and making the model available to multiple participants.

Description and Results

As shown in Figure 2-6-6, this case uses the U-Net topology of deep Convolutional Neural Network (CNN), which takes a single-channel image as input and outputs an equivalent binary mask that assigns a classification label to each pixel. The network simulates the architecture of the auto-encoder that is capable of max-pooling, capturing context with contracting path and achieving localized extension path via up-sampling. Unlike standard auto-encoders, each feature map in the extension path is associated with a corresponding feature map in the contracted path by using the skip connection, which allows the model to acquire more downstream feature maps with spatial information through a smaller receptive field. Simply speaking, this allows the network to consider features at different spatial scales.

U-Net is now one of the standard deep learning topologies for medical image segmentation and plays a huge role in workloads such as neural ultrasound image segmentation and lung CT scan image segmentation. The model was used for each of the federated learning validation tests in this case, with the Dropout parameter set to 0.2 and the up-sampling set to true. For more information on how to optimize the U-Net segmentation network, please refer to the previous article.

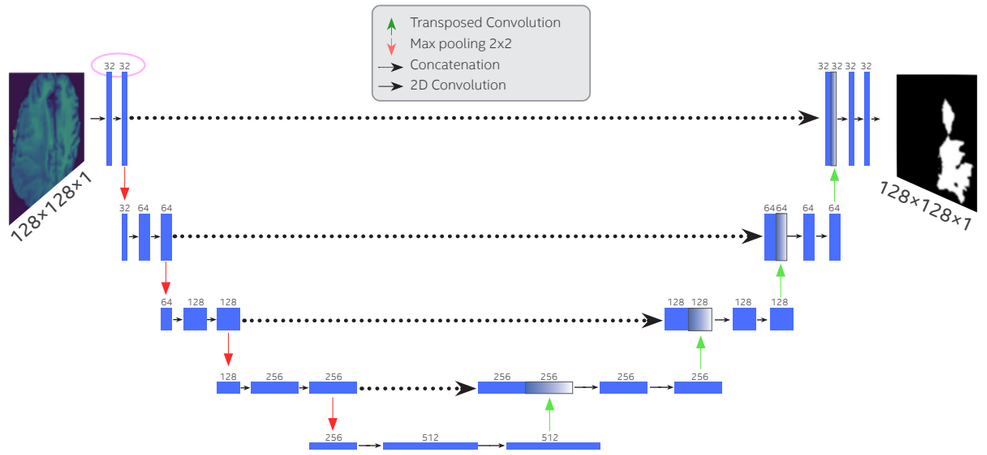


Figure 2-6-6 U-Net Topology for the BraTS Federated Learning Solution[60]

[59] For details, please refer to http://www.miccai.org/

[60] Image cited from Multi-Institutional Deep Learning Modeling Without Sharing Patient Data: A Feasibility Study on Brain Tumor Segmentation, Micah J Sheller, G Anthony Reina, Brandon Edwards, Jason Martin, Spyridon Bakas, https://arxiv.org/pdf/1810.04304v1.pdf

The BraTS 2018 training dataset[61] was used in this solution, which contains multi-modal scanning Magnetic Resonance Imaging (MRI) of the brain from patients diagnosed with glioma in multiple healthcare organizations. The radiographic abnormal areas of each brain scan have been manually annotated with three distinct labels corresponding to peritumoral edema/tissue invasion, non-enhancing/solid and necrotic/cystic tumor core, and enhancing tumor. Since this case was designed to evaluate the performance of federated learning in clinical image segmentation, only the volume of tumors annotated with the above three labels was of interest. The case also selected collaborative learning methods such as IIL and CIIL as a control group.

The federated learning architecture is shown in Figure 2-6-7, where none of the participants need to share their respective data but instead train the shared model locally and only send model updates to the aggregator. The aggregator consolidates the updates and sends the new shared parameters to various participants for further training (which can be iterated) or application. The consolidated shared parameters are equivalent to the weighted average of the updates provided by each participant, and the weight of a particular participant is given as the fraction of the total data instances residing in that participant. This iterative process of local training, update consolidating, and distribution of new parameters is called a collaborative round. The solution evaluates the impact of different numbers of participants, and different EpR, on the performance of the final AI application. For more details on the process of federated learning solution, see relevant description of "Typical solution based on Intel® SGX" on page 72.

The performance of BraTS image segmentation can be evaluated by the **Dice Coefficient (DC)** value, which reflects the intersection ratio between predicted and actual unions and can be defined as:

$$DC = \frac{2|P \cap T|}{|P| + |T|}$$

where **P** and **T** are the Masks for Prediction and Ground Truth (GT), respectively. The benchmark value in the solution are obtained from the U-Net topology, trained on Data that has been fully Shared (Data-Sharing). Its verified peak accuracy **DC = 0.862** (optimal value). As shown in Figures 2-6-8, the top side shows the change in DC values for various collaborative learning methods in each collaborative round. It can be found that the DC values for the federated learning approach are the most stable ones and close to the optimal values obtained under the full data sharing approach, while the DC values for IIL and CIIL approaches are more volatile. The bottom side shows the verified DC values for various collaborative learning methods after each complete training, and the DC values for the federated learning approach are also close to the optimal value and very stable.

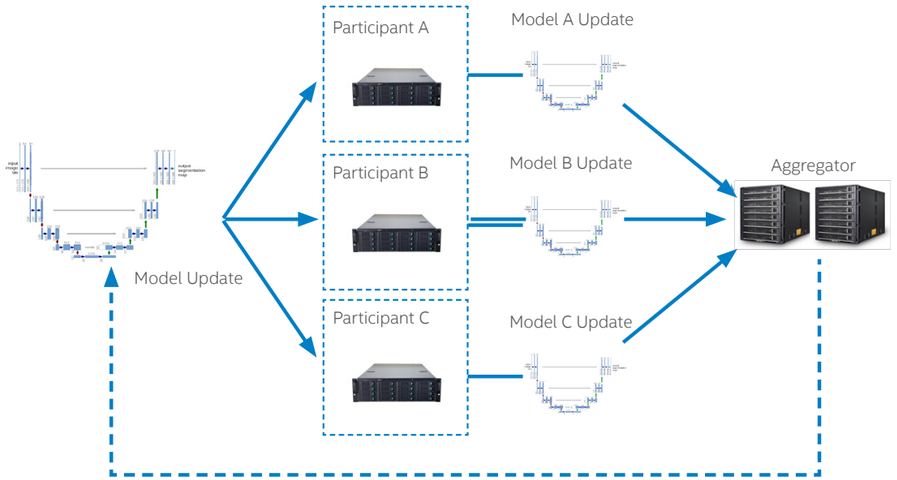


Figure 2-6-7 Architecture of Federated Learning Solution for BraTS

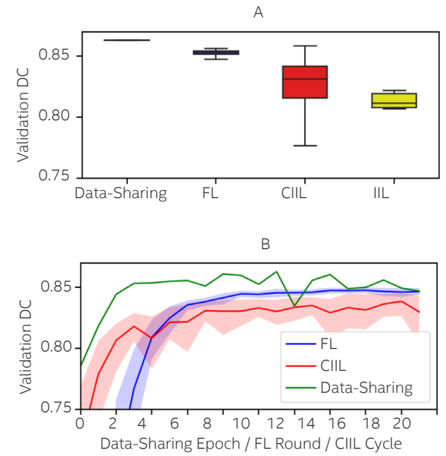


Figure 2-6-8 Performance of Federated Learning Compared to Other Learning Methods[62]

The results of the validation tests show that, in healthcare organizations, the performance of federated learning approach can reach 99% of the performance of the full data sharing approach[63], even when using unbalanced datasets. Undoubtedly, by introducing a federated learning approach, healthcare organizations can more effectively improve and enhance the performance of computer-assisted analysis and diagnostic systems, thereby facilitating the development of precision medicine, as well as effectively addressing a range of security, privacy, or data ownership issues that arise from data sharing.

For details about the case, please refer to: Multi-Institutional Deep Learning Modeling Without Sharing Patient Data: A Feasibility Study on Brain Tumor Segmentation, Micah J Sheller, G Anthony Reina, Brandon Edwards, Jason Martin, Spyridon Bakas
https://arxiv.org/pdf/1810.04304v1.pdf

Conclusion

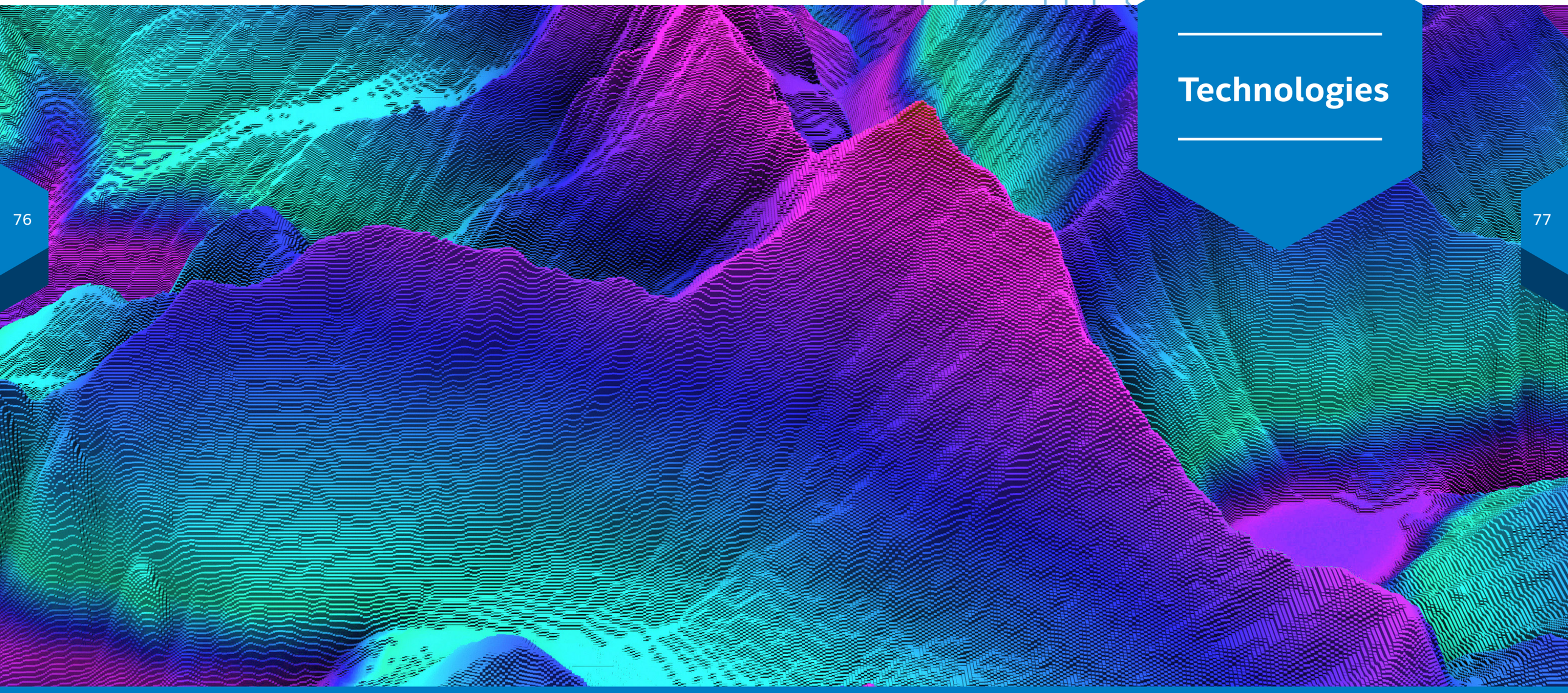
As an important "fuel" for the development of medical AI applications, more high-quality medical data can undoubtedly enhance the performance of AI applications, but how to address the data security and privacy issues has always been a challenge for the healthcare industry when promoting AI development. The federated learning approach, however, has proven to be a good solution to this challenge.

Now, Intel is working with many medical and research institutions to leverage advanced hardware and software products such as Intel® SGX and Intel® processors and use federated learning approach, in order to effectively address the lack of training data faced by AI training in medical institutions while ensuring data security and reliability, thereby further promoting the rapid development of medical AI applications.

[61] The dataset is cited from Menze, B.H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J.,Burren, Y., Porz, N., Slotboom, J., Wiest, R., Lanczi, L., Gerstner, E., Weber, M.A.,Arbel, T., Avants, B.B., Ayache, N., Buendia, P., Collins, D.L., Cordier, N., Corso,J.J., Criminisi, A., Das, T., Delingette, H., , D., Durst, C.R., Dojat, M., Doyle,S., Festa, J., Forbes, F., Geremia, E., Glocker, B., Golland, P., Guo, X., Hamamci,A., Iftekharuddin, K.M., Jena, R., John, N.M., Konukoglu, E., Lashkari, D., Mariz,J.A., Meier, R., Pereira, S., Precup, D., Price, S.J., Raviv, T.R., Reza, S.M.S.,Ryan, M., Sarikaya, D., Schwartz, L., Shin, H.C., Shotton, J., Silva, C.A., Sousa,N., Subbanna, N.K., Szekely, G., Taylor, T.J., Thomas, O.M., Tustison, N.J., Unal,G., Vasseur, F., Wintermark, M., Ye, D.H., Zhao, L., Zhao, B., Zikic, D., Prastawa,M., Reyes, M., Leemput, K.V.: The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS). IEEE Transactions on Medical Imaging 34(10), 1993-2024(2015)、Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J.S., Freymann,J.B., Farahani, K., Davatzikos, C.: Advancing The Cancer Genome Atlas gliomaMRI collections with expert segmentation labels and radiomic features. Nature Scientific Data 4, 170117 (2017) https://doi.org/10.1038/sdata.2017.117、 Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J., Freymann, J.,Davatzikos, C.: Segmentation Labels and Radiomic Features for the Pre-operativeScans of the TCGA-GBM collection. In: The Cancer Imaging Archive, (2017) 以及 Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J., Freymann, J.,Davatzikos, C.: Segmentation Labels and Radiomic Features for the Pre-operativeScans of the TCGA-LGG collection. In: The Cancer Imaging Archive, (2017)

[62] Image cited from Multi-Institutional Deep Learning Modeling Without Sharing Patient Data: A Feasibility Study on Brain Tumor Segmentation, Micah J Sheller, G Anthony Reina, Brandon Edwards, Jason Martin, Spyridon Bakas, https://arxiv.org/pdf/1810.04304v1.pdf

[63] Data cited from Multi-Institutional Deep Learning Modeling Without Sharing Patient Data: A Feasibility Study on Brain Tumor Segmentation, Micah J Sheller, G Anthony Reina, Brandon Edwards, Jason Martin, Spyridon Bakas, https://arxiv.org/pdf/1810.04304v1.pdf



Technologies

2nd Generation Intel® Xeon® Scalable Processor

The 2nd Generation Intel® Xeon® Scalable Processor is designed specifically for data center modernization, and it provides 25%-35% higher performance[1] than its previous generation. With many new features, improved flexibility and security, and enhanced memory performance, it facilitates users to improve the operational efficiency of various infrastructures, enterprise applications and technical computing, and deliver agile services with enhanced performance and more valuable capabilities that lead to improved Total Cost of Ownership (TCO) and higher productivity.

The Intel® Xeon® Gold 6200 processor series, especially the mainstream Intel® Xeon® Gold 6248 processor, Intel® Xeon® Gold 6240 processor and Intel® Xeon® Gold 6230 processor, are the mainstay of the Intel® Xeon® Scalable Processor platform. With support for the higher memory speeds, enhanced memory capacity, and four-socket scalability, it delivers significant improvement in performance, advanced reliability, and hardware-enhanced security. It is optimized for demanding mainstream data center, multi-cloud compute, and network and storage workloads, and it is suitable for more complex and diversified application scenarios. In addition, the Intel® Xeon® Gold 6200 series supports dual FMA channels for the first time, which means that the FMA performance has improved by 2 times[2].

Moreover, the 2nd Generation Intel® Xeon® Scalable Processor integrates Deep Learning Boost (Vector Neural Network Instruction, VNNI), expands Intel® AVX-512, and empowers the platform with more and stronger AI capabilities to accelerate AI and deep learning inference and optimize workloads. This gives it a CPU architecture that integrates AI acceleration capabilities. Based on this architecture, most inference workloads are integrated into workloads or applications, providing users with the advantages of accelerated performance and higher flexibility. It provides the seamless performance foundation for the data centric era from the multi-cloud to intelligent edge, and back, as well as AI development and application.

As the "core" of the next generation Xeon® scalable platform, the 2nd Generation Intel® Xeon® Scalable Processor supports the new category of Intel® Optane™ persistent memory. When used with the 2nd generation Intel® Xeon® Gold and Platinum processors, Intel® Optane™ persistent memory complements DRAM to significantly improve system performance and accelerate workload processing and service delivery. (For details, please refer to the section **"Intel® Optane™ Persistent Memory"**).

Features:

- **Higher Per-Core Performance:** Up to 56 cores (9200 series) and up to 28 cores (8200 series), delivering high performance and scalability for compute-intensive workloads across compute, storage, and network usages.
- **Intel® DL Boost with VNNI:** It brings enhanced artificial intelligence inference performance on CPU. With up to 30X performance improvement over the previous generation[3], it helps to deliver AI readiness and application from the data center to the edge.
- **Industry-leading memory and storage support, greater memory bandwidth/capacity:** Support for Intel® Optane™ persistent memory, supporting up to 36 TB of system memory capacity when combining with traditional DRAM; 50% increased memory bandwidth and capacity[4]. Support for six memory channels and up to 4 TB of DDR4 memory, per socket, with speeds up to 2933 MT/s (1 DPC). It also supports Intel® Optane™ SSDs and QLC 3D NAND SSDs. For data-intensive workloads, breakthrough memory and storage memory innovations can significantly improve their efficiency and performance.
- **Intel® Infrastructure Management Technologies (Intel® IMT):** This framework for resource management combines multiple Intel capabilities that effectively support platform-level detection, reporting and configuration.
- **Intel® Security Libraries for Data Center (Intel® SecL-DC):** This set of software libraries and components enables Intel hardware-based security features.

As an innovation of the Xeon® platform, disruptive by design, the 2nd Generation Intel® Xeon® Scalable Processor sets a new level of platform convergence and capabilities across compute, storage, memory, network, and security.

[1] https://www.intel.com/content/www/us/en/technology-provider/products-and-solutions/xeon-scalable-family/2gen-data-centric-computing-article.html
[2, 3, 4] https://www.intel.com/content/www/us/en/products/docs/processors/xeon/2nd-gen-xeon-scalable-processors-brief.html

2nd Generation Intel® Xeon® Scalable Processor for AI Application

*Supported on select processors only.

	Intel® Xeon® Gold Processor (6200 Series)	Intel® Xeon® Gold Processor (6200 Series)						Intel® Xeon® Platinum Processor (8200 Series)	Intel® Xeon® Platinum Processor (9200 Series)
		Intel® Xeon® Gold 6230 Processor	Intel® Xeon® Gold 6230R Processor	Intel® Xeon® Gold 6240 Processor	Intel® Xeon® Gold 6240R Processor	Intel® Xeon® Gold 6248 Processor	Intel® Xeon® Gold 6248R Processor		
Pervasive Performance and Security									
Highest Core Count Supported	24 cores	20 cores	26 cores	18 cores	24 cores	20 cores	24 cores	28 cores	56 cores
Highest Supported Frequency	4.4 GHz	3.90 GHz	4.00 GHz	3.90 GHz	4.00 GHz	3.90 GHz	4.00 GHz	4.0 GHz	3.8 GHz
Number of CPU Sockets Supported	Up to 4 sockets	Up to 4 sockets	Up to 2 sockets	Up to 4 sockets	Up to 2 sockets	Up to 4 sockets	Up to 2 sockets	Up to 8 sockets	Up to 2 sockets
Intel® Ultra Path Interconnect (UPI)	3	3	2	3	2	3	2	3	4
Intel® UPI Speed	10.4 GT/s	10.4 GT/s	10.4 GT/s	10.4 GT/s	10.4 GT/s	10.4 GT/s	10.4 GT/s	10.4 GT/s	10.4 GT/s
Intel® Advanced Vector Extensions 512 (Intel® AVX-512)	2 FMA	2 FMA	2 FMA	2 FMA	2 FMA	2 FMA	2 FMA	2 FMA	2 FMA
Hightest Memory Speed Support (DDR4)	2933 MT/s	2933 MT/s	2933 MT/s	2933 MT/s	2933 MT/s	2933 MT/s	2933 MT/s	2933 MT/s	2933 MT/s
Highest Memory Capacity Supported Per Socket*	1 TB, 2 TB, 4.5 TB	1 TB	1 TB	1 TB	1 TB	1 TB	1 TB	1 TB, 2 TB, 4.5 TB	3.0 TB
16 Gb DDR4 DIMM Support	●	●	●	●	●	●	●	●	●
Intel® Deep Learning Boost (Intel® DL Boost) with Vector Neural Network Instruction (VNNI)	●	●	●	●	●	●	●	●	●
Intel® Optane™ Persistent Memory Module Support*	●	●	●	●	●	●	●	●	
Intel® Omni-Path Architecture (Discrete PCIe* card)	●	●	●	●	●	●	●	●	●
Intel® QuickAssist Technology (Integrated in chipset)	●	●	●	●	●	●	●	●	
Intel® QuickAssist Technology (Discrete PCIe* Card)	●	●	●	●	●	●	●	●	●
Intel® Optane™ SSDs	●	●	●	●	●	●	●	●	●
Intel® SSD Data Center Family (3D NAND)	●	●	●	●	●	●	●	●	●
PCIe 3.0	●	●	●	●	●	●	●	●	●
Intel® QuickData Technology (CBDMA)	●	●	●	●	●	●	●	●	●
Non-Transparent Bridge (NTB)	●	●	●	●	●	●	●	●	●
Intel® Turbo Boost Technology 2.0	●	●	●	●	●	●	●	●	●
Intel® Hyper-Threading Technology (Intel® HT Technology)	●	●	●	●	●	●	●	●	●
Node Controller Support	●	●	●	●	●	●	●	●	●

*Supported on select processors only.

For more information on the 2nd Generation Intel® Xeon® Scalable Processor, please visit:

https://www.intel.com/content/www/us/en/products/processors/xeon/scalable.html

Intel® Optane™ Persistent Memory

Intel's breakthrough product, Intel® Optane™ persistent memory, is disrupting the traditional memory-storage hierarchy by creating a new tier to fill the memory-storage gap, and providing a large persistent memory hierarchy at a reasonable price, which can provide greater performance, efficiency, and affordability for memory-intensive workloads, virtual machine density and fast storage. Then it facilitates users to accelerate IT transformation through faster analysis, cloud services, artificial intelligence training and inference, and next generation communication services than ever before to meet the needs of the data era.

Intel® Optane™ persistent memory combines cost, capacity, non-volatile and performance features, and a single module provides 128/256/512 GiB options available and is compatible with DDR4 socket. When applied on the platform based on the 2nd Generation Intel® Xeon® Scalable Processor together with the traditional DDR4 DRAM memory, it can achieve the capacity of up to 24TiB on the eight socket system at lower cost (each socket providing up to 3TiB of Optane persistent memory), so as to facilitate users to load datasets on the memory system close to the processor that are far larger than previous datasets, meet the application load needs with demanding requirements on the large memory including memory database, larger-scale dataset analysis, and AI inference and iteration, etc., and enable more data processing and analysis to be performed in real time.

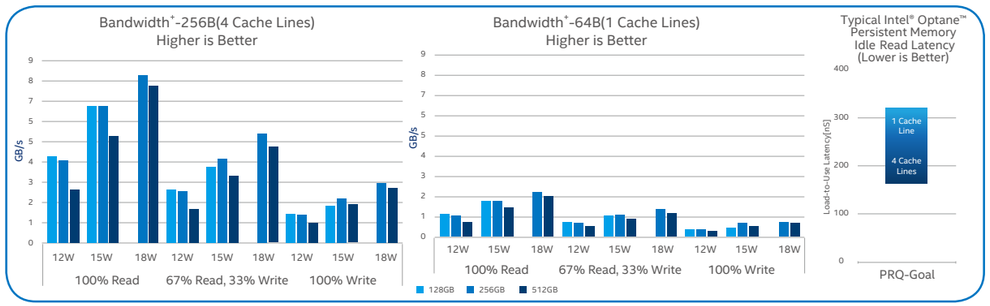
Based on the combination of memory and storage features, Optane™ persistent memory has two operating modes: Memory Mode and App Direct Mode. With distinct operating modes, customers have the flexibility to significantly improve system performance across multiple workloads.

Memory Mode. Memory Mode is great for large memory capacity and does not require application changes. In Memory Mode, the CPU memory controller uses Intel® Optane™ persistent memory as addressable main memory to extend the amount of available volatile memory visible to the Operating System, while using the DRAM as cache. This can directly facilitate virtualization and container technology because it can increase the density of virtual machines or containers on a single physical server at lower cost, or provide larger memory capacity for each virtual machine and container without rewriting software.

App Direct Mode. In App Direct Mode, the operating system treats DRAM and Intel® Optane™ persistent memory as two separate pools of memory. It is byte addressable like memory, persistent like storage. The power of persistent memory adds business resilience to systems with faster restart times and lower cost because data is retained even during maintenance or power cycles.

Mixed Mode. It is a subset of App Direct and can be provisioned so that some of Intel® Optane™ persistent memory is in Memory Mode and the remaining is in App Direct Mode, thus meeting the needs of workloads or application environments with dual demands.

Intel® Optane™ persistent memory is launched with other Intel® Xeon® scalable processor platform products and optimized for the 2nd Generation Intel® Xeon® Scalable Processor. Its unique application advantages have attracted many ISV partners to start tuning their related software from the very beginning of its launch. Meanwhile, there are also cloud infrastructure and data analysis users who have also introduced Intel® Optane™ persistent memory. These partners and users have initially witnessed the application value of Intel® Optane™ persistent memory in multiple modes in the corresponding application tuning and application practice.



Product Family	Intel® Optane™ Persistent Memory					
Form Factor	Persistent Memory Module (PMM)					
PMem SKU[1]	128 GiB		256 GiB		512 GiB	
User Capacity	126.4 GiB[8]		252.4 GiB[8]		502.5 GiB[8]	
MOQ	4	50	4	50	4	50
MM#	999AVV	999AVW	999AVX	999AVZ	999AW1	999AW2
Product Code	NMA1XXD128GPSU4	NMA1XXD128GPSUF	NMA1XXD256GPSU4	NMA1XXD256GPSUF	NMA1XXD512GPSU4	NMA1XXD512GPSUF
Model String	NMA1XXD128GPS		NMA1XXD256GPS		NMA1XXD512GPS	
TECHNOLOGY	Intel® Optane™ Technology					
Limited Warranty	5 years					
Annual Failure Rate (AFR)	≤ 0.44					
Endurance 100% Write 15W 256B	292 PBW		363 PBW		300 PBW	
Endurance 100% Write 15W 256B	91 PBW		91 PBW		75 PBW	
Endurance 100% Read 15W 64B	6.8 GB/s		6.8 GB/s		5.3 GB/s	
Endurance 100% Write 15W 256B	1.85 GB/s		2.3 GB/s		1.89 GB/s	
Endurance 100% Read 15W 64B	1.7 GB/s		1.75 GB/s		1.4 GB/s	
Endurance 100% Write 15W 64B	0.45 GB/s		0.58 GB/s		0.47 GB/s	
DDR Frequency	2666, 2400, 2133, 1866 MT/s					
Maximum Thermal Design Power (TDP)	15W		18W			
Temperature (TJMAX)	≤ 84℃ (85℃ shutdown, 83℃ default) media temperature					
Temperature (Ambient Temperature)	10W: 54℃ @ 2.4m/s					
Temperature (Ambient Temperature)	12W: 49℃ @ 2.4m/s					
Temperature (Ambient Temperature)	15W: 44℃ @ 2.7m/s					
Temperature (Ambient Temperature)	N/A		18W: 40℃ @ 3.7m/s			

Note: [1]GiB = 2^{30}; GB = 10^{9}

For more information, please refer to
https://www.intel.com/content/www/us/en/products/memory-storage/optane-dc-persistent-memory.html

Intel® Optane™ SSDs and Intel® SSDs with Intel® QLC 3D NAND Technology

Intel® Optane™ SSDs and Intel® SSDs with Intel® QLC 3D NAND Technology, with their innovative storage hierarchy, helps to build a future-ready data center to accelerate changes and achieve great leaps.

As a high-end member of Intel's SSD product line, Intel® Optane™ SSDs feature innovative 3D XPoint™ storage media and incorporate advanced system memory controllers, interface hardware, and software technologies, delivering low latency and high stability. It is designed to alleviate data center storage bottlenecks and allow for bigger, more affordable datasets, thus accelerating applications, reducing transaction costs for latency-sensitive workloads, and improving data center TCO. Intel® Optane™ SSDs, with more comprehensive, better and more balanced IT infrastructure capability, can undoubtedly bring higher efficiency to data-intensive AI model training and inference. Take Intel® Optane™ SSD DC P4800X as an example. It offers up to 550,000 IOPS of random read/write capacity and less than 10 ms of read/write latency, enabling the solution to perform more effectively in multi-user and high-concurrency scenarios. Meanwhile, its Drive Writes Per Day (DWPD) is as high as 60, which is much higher than NAND SSDs, so it can give the storage system a longer lifespan and ensure better value.

Intel® SSDs use breakthrough, trusted 3D NAND technology to improve storage, thus providing a cost-effective replacement for traditional Hard Disk Drives (HDD), and helping customers accelerate user experiences, improve the performance of applications and services and reduce TCO. As a "new force" in SSDs, the Intel® SSD D5-P4320 series rely on Intel's leading 64-layer 3D NAND technology to enable a single QLC SSD disk capacity of up to 7.68 TB (TeraBytes) in order to adequately fulfill the storage requirements of infrastructure users such as data centers for massive data. It also has a random read IOPS of up to 427,000, and when paired with the 2nd Generation Intel® Xeon® Scalable Processor, it is especially suitable in terms of meeting "Write Once, Read Many" (WORM) performance requirements in application scenarios, such as AI training, providing a storage framework with high efficiency, high stability and low energy consumption to support complex and diverse workloads.

To find out more, visit:

- https://www.intel.cn/content/www/cn/zh/products/memory-storage/optane-memory/optane-memory-h10-solid-state-storage.html
- https://www.intel.cn/content/www/cn/zh/products/memory-storage/solid-state-drives/data-center-ssds.html

Unified Open Source Big Data Analytics + AI Platform Analytics Zoo

Analytics Zoo is a unified big data analytics + AI open source platform. It is designed for users to develop big data-based and end-to-end deep learning applications.

Analytics Zoo allows users to implement TensorFlow, Keras, Pytorch and BigDL, as well as other frameworks that may need to be supported in the future on top of Apache Spark/Flink and Ray, seamlessly integrate them into an pipeline, and transparently scale out these models to large data clusters with hundreds or thousands of nodes for distributed training or inference, thereby further simplifying the development of AI solutions without additional dedicated infrastructure.

In order to improve computing performance, Analytics Zoo integrates various software libraries, such as Intel® MKL and Intel® MKL-DNN. In terms of hardware, it is based on the Intel® Xeon® processor platform and fully releases the integrated vector and deep learning instructions of the 2nd Generation Intel® Xeon® Scalable Processor, which can greatly improve the training and inference speed.

The benefits of integrating data storage and processing pipelines into a unified infrastructure without moving data are obvious - it can not only improve development and deployment efficiency and scalability, reduce hardware management and developers' time to learn new languages, improve development and deployment efficiency, resource utilization and flexibility, but also reduce total cost of ownership without affecting computing efficiency and performance. Developers just need to make full use of the rich features and functions provided by Analytics Zoo and various analytics and AI tools when scaling up their AI solutions, so as to achieve the efficient integration, deployment and application of big data analytics and AI.

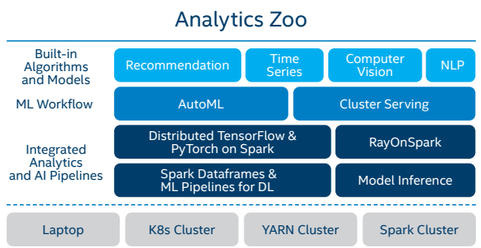
Among the many AI frameworks supported by Analytics Zoo platform, BigDL is an open source platform developed by Intel itself. BigDL is a distributed deep learning framework based on Apache Spark, which can run seamlessly and directly on top of existing Apache Spark and Hadoop clusters without any modification to the clusters. Based on BigDL, developers can write deep learning applications as Scala or Python programs and take advantage of the power of scalable Spark clusters to promote the integration of big data analytics and AI. Users who have been familiar with and enabled BigDL in the past few years can invoke BigDL directly through Analytics Zoo for seamless switching.

Technical Features of Analytics Zoo:

- **End-to-end pipeline, applying AI models (TensorFlow, PyTorch, OpenVINO™ toolkit, etc.) to distributed big data:**
  - Distributed training and prediction with Spark code for TensorFlow or PyTorch
  - Native deep learning support (TensorFlow/Keras/PyTorch/ BigDL) in the Spark ML pipeline
  - With RayOnSpark, run Ray programs directly on big data clusters
  - Provide Plain Java/Python APIs (TensorFlow/PyTorch/ BigDL/ OpenVINO™) to serve Model Inference
- **Highly abstract ML workflows to automate machine learning tasks**
  - Cluster Serving for automated distributed (TensorFlow/PyTorch/Caffe/ OpenVINO™) model inference
  - Extensible AutoML for time series-based data analysis and prediction
- **Built-in models for recommender systems, time series analysis, computer vision and natural language processing applications**

Why Analytics Zoo:

- AI models (e.g., TensorFlow, Keras, PyTorch, BigDL, OpenVINO™ toolkit) can be easily applied to distributed big data.
- AI applications can be scaled transparently from a single laptop to large clusters with "zero" code changes.
- AI pipelines can be deployed to existing YARN or K8S clusters without making any changes to the cluster.
- The process of applying machine learning (e.g. feature engineering, hyper-parameter tuning, model selection, distributed inference) can be automated.



To find out more, visit:
https://software.intel.com/zh-cn/blogs/2018/09/10/analytics-zoo-unifying-analytics-ai-for-apache-spark?elq_cid=4287274&erpm_id=7282583

Intel® Data Analytics Acceleration Library

As a branch of artificial intelligence, machine learning is now gaining tremendous attention and advanced analytics based on machine learning is becoming increasingly popular due to its ability to facilitate IT staff, data scientists, various business teams and their organizations quickly unlock advantages over other analytics. And machine learning offers many new commercial and open source solutions, providing a rich ecosystem for developers. In addition, developers can choose from a variety of open source machine learning libraries such as Scikit-learn, Cloudera and Spark MLlib.

Intel® Data Analytics Acceleration Library (Intel® DAAL)

To facilitate its industry users deploy machine learning, Intel also offers a high-performance and systematic solution that covers a rich set of resources including processors, optimized software and developer support, and a robust ecosystem.

Machine learning requires robust computing power. Intel® Xeon® processors provide a scalable benchmark, which is specifically designed to meet the highly parallel workloads unique to machine learning and its memory and architecture (networking) requirements. In an Intel test, the processor reduced system training time by a factor of 50[1].

In addition, Intel provides software support, including:

- Libraries, languages, and building blocks optimized for Intel® Xeon® processors, oneDNN and **Intel® Data Analytics Acceleration Library (Intel® DAAL)**, and Intel® Distribution for Python.
- Optimization frameworks that simplify development, including Apache Spark, Caffe, Torch, and TensorFlow. Intel supports both open source and commercial software, enabling users to quickly take advantage of the latest processor and system features available on the market.

Intel® DAAL is a suite of end-to-end software solutions designed to facilitate data scientists and analysts quickly build everything from data pre-processing, to data feature engineering, data modeling and deployment. It provides various data analytics needed to develop machine learning and analytics as well as high-performance building blocks required by algorithms. Classical machine learning algorithms such as linear regression, logistics regression, LASSO, AdaBoost, Bayesian classifiers, support vector machines, K nearest neighbors, Kmeans clustering, DBSCAN clustering, various decision trees, random forests, Gradient Boosting are now supported. These algorithms are highly optimized to achieve high performance on Intel® processors. For example, a leading big data analytics technology and services provider in China has used these resources to improve the performance of data mining algorithms by 3X to 14X[2].

To make it easier for developers to use Intel® DAAL in machine learning applications in Intel-based environments, Intel has open-sourced the entire project (https://github.com/intel/daal), and provides full-memory, streaming and distributed algorithm support for different big data scenarios. For example, DAAL Kmeans can be well combined with Spark to perform multi-node clustering on a Spark cluster. In addition, Intel® DAAL provides interfaces to C++, Java, and Python.

DAAL4py

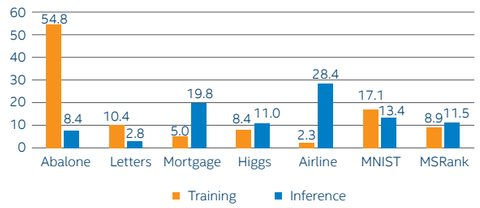
To deliver better support for the most widely used Python application Scikitlearn, Intel® DAAL provides a very simple Python interface DAAL4py (open source resources at https://github.com/IntelPython/daal4py), which works seamlessly with Scikitlearn and enables underlying algorithmic acceleration for machine learning. Developers do not need to modify the Scikitlearn code to take advantage of automatic vectorization and multi-threading. DAAL4py currently supports the following algorithms in Scikitlearn:

- sklearn. linear regression, sklearn. ridge regression, logistics regression
- PCA
- KMeans
- pairwise_distance
- SVC (SVM classification)

Intel-optimized XGBoost

XGBoost is a machine learning open source library based on progressive Gradient Boosting, which is widely used in a variety of classification and decision making businesses. To further enhance its performance, Intel has optimized and open-sourced the codebase[3] and the latest optimizations have been integrated into XGBoost 1.0 and later versions. Compared to XGBoost version 0.9, the new version offers a 2X-54X performance increase.[4]

Gradient Boosting Performance (Higher is better) Intel® DAAL 2020 vs DMLC XGBoost 0.9 Speed-Up



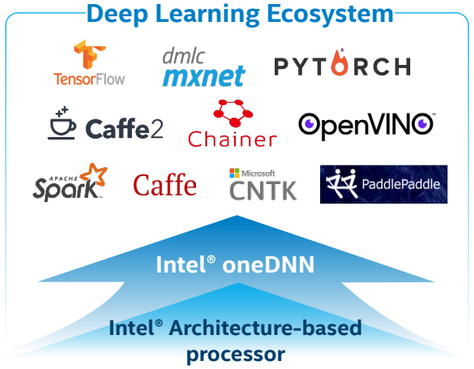
For more information, visit: https://software.intel.com/en-us/daal

[1], [2] https://software.intel.com/zh-cn/articles/meritdata-speeds-up-a-big-data-platform
[3] Performance optimizations for Intel CPUs : https://github.com/dmlc/xgboost/pull/3957/files
[4] https://software.intel.com/daal

Intel® Deep Neural Network Library (oneDNN)

The Intel® Deep Neural Network Library (formerly the Intel® Math Kernel Library for Deep Neural Networks, Intel® MKL-DNN) is an open source, performance enhancement library for deep-learning applications, and a foundation library created by Intel to facilitate developers make full use of Intel® architecture to promote the research and application of deep learning. (Source Code Address: https://github.com/intel/mkl-dnn)

oneDNN, as a performance enhancement library specifically designed for accelerating deep learning framework on Intel® architecture, includes highly vectorized and threaded building blocks for implementing deep neural networks with C and C++ interfaces, and has a broad ecosystem of deep learning research, development and applications. It currently supports: a rich set of deep learning software products such as TensorFlow, PyTorch, MXNet, Caffe, Spark BigDL, OpenVINO™ toolkit.



In order to effectively accelerate deep learning model on Intel® architecture, and improve the efficiency of other performance-sensitive applications in various neural networks, oneDNN provides a number of optimized deep learning operation primitives that can be used in different deep learning frameworks to ensure the efficient implementation of common building blocks. These modules include:

- **Matrix multiplication and convolution:** 1D/2D/3D, Winograd 2D
- **RNN primitives;**
- **Inner product;**
- **Pooling:** maximum, minimum, average;
- **Normalization:** Local Response Normalization across channels (LRN), batch normalization;
- **Activation:** Rectified Linear Unit (ReLU);
- **Data manipulation:** multi-dimensional transposition (conversion), split, concat, sum and scale.

These efficient function blocks can be applied to a wide range of deep learning models, such as:

Application type	Topologies
Image Recognition	AlexNet, VGG, GoogleNet, ResNet, MobileNet
Image Segmentation	FCN, SegNet, MaskRCNN, U-Net
Volumetric Segmentation	3D-Unet
Object Detection	SSD, Faster R-CNN, Yolo
Machine Translation	GNMT
Speech Recognition	DeepSpeech, WaveNet
Adversarial Networks	DCGAN, 3DGAN
Reinforcement Learning	A3C

In order to greatly improve the performance of deep learning on CPU, Intel has also cooperated with many open source communities to integrate this library into various deep learning frameworks. For example, as early as 2016, Caffe optimized by oneDNN achieved a performance improvement of up to 10X over the original Caffe using Intel® Xeon® Processor E5-2697 v3[1]. In 2019, the optimized ResNet-50 also achieves the leading performance of 7,736 images per second on Intel® Xeon® Platinum 9282 Processor[2].

oneDNN has now become the basic configuration of many deep learning frameworks running on CPU. Developers can directly benefit from the performance improvement brought by oneDNN in the installation and application of the deep learning framework.

To find out more, visit:

- https://software.intel.com/zh-cn/articles/intel-mkl-dnn-part-1-library-overview-and-installation
- https://software.intel.com/zh-cn/articles/introducing-dnn-primitives-in-intelr-mkl

[1] https://software.intel.com/es-es/node/604830?language=en
[2] https://www.intel.com/content/dam/www/public/us/en/images/diagrams/rwd/xeon-scalable-max-inference-rwd.png

Intel® Optimization for Caffe

Intel® Optimization for Caffe was integrated with the then current release of Intel® MKL from the original version, and it is specially optimized for Intel® AVX 2 and Intel® AVX-512 which were integrated with Intel® Xeon® processors at that time. It is fully compatible with BVLC Cafe, and incorporates all the advantages of BVLC Cafe. With processor optimization feature, it shows excellent performance on Intel® architecture processors, and supports multi-node distributed program training.

Intel® Optimization for Caffe supports a complete set of Post-training quantization options and has been implemented in a large number of CNN models. Especially in the platform based on the 2nd Generation Intel® Xeon® Scalable Processor (see the processor section), the integrated Intel® Deep Learning Boost (VNNI instruction set) with optimization support for INT8 allows for accelerating the inference speed of multiple deep learning models when using INT8 up to 2-4 times that when using FP32 (see the following figure)[1] without affecting the prediction accuracy, thus greatly improving the working efficiency of users' deep learning applications.

To find out more, visit:

- https://software.intel.com/en-us/articles/caffe-optimized-for-intel-architecture-applying-modern-code-techniques
- https://software.intel.com/zh-cn/videos/what-is-intel-optimization-for-caffe

Optimized Deep Learning Frameworks and Toolkits

Benefits of ResNet-50 with Intel® Deep Learning Boost

2nd Generation Intel® Xeon® Platinum 8280 Processor vs Intel® Xeon® Platinum 8180 Processor

Intel® Xeon® Scalable Processor	2nd Generation Intel® Xeon® Scalable Processor	mxnet	PYTORCH	TensorFlow	Caffe	OpenVINO
FP32	INT8 W/ Intel® DL Boost	3.0x	3.7x	3.9x	4.0x	3.9x
INT8	INT8 W/ Intel® DL Boost	1.8x	2.1x	1.8x	2.3x	1.9x

[1] For configuration details, see: https://www.intel.com/content/www/us/en/benchmarks/server/xeon-scalable/xeon-scalable-artificial-intelligence.html

Intel® Optimization for TensorFlow

Intel® Optimization for TensorFlow is an optimized version launched by Intel to deal with the performance challenges of running deep learning models on CPU, making sure that deep learning workloads can run efficiently with Intel® MKL-DNN basic computing units under all conditions.

In order to significantly improve performance, Intel continues to optimize TensorFlow in other ways.

Graph Optimizations

Intel has introduced a number of graph optimization passes to replace default TensorFlow operations with Intel optimized versions when running on CPU. This ensures that users can run their existing Python programs and obtain the performance gains without changes to their neural network model. Meanwhile, it can eliminate unnecessary and costly data layout conversions, fuse multiple operations together to enable efficient cache reuse on CPU, and handle intermediate states that allow for faster backpropagation.

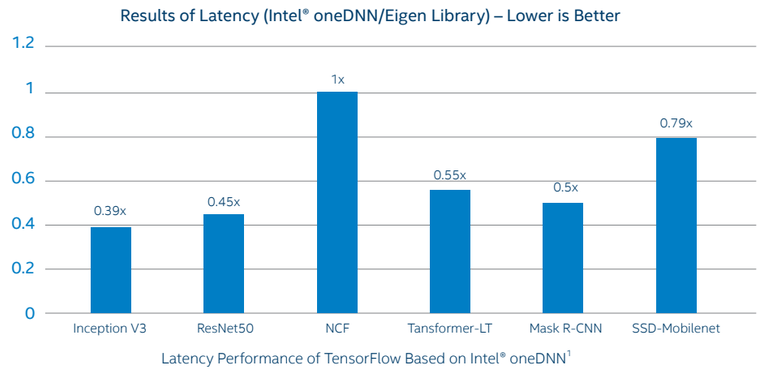
These graph optimizations enable greater performance without introducing any additional burden on TensorFlow programmers. In addition, data layout optimization is a key performance optimization. Previously, Intel inserts a data layout conversion operation from TensorFlow's native format to an internal format, performs the operation on CPU and converts operation output back to the TensorFlow format, but these conversions introduce a performance overhead. Now, the sub-graphs that can be entirely executed using Intel® MKL optimized operations can eliminate the conversions within the operations in the sub-graph. Automatically inserted conversion nodes take care of data layout conversions at the boundaries of the sub-graph. Another key optimization is the fusion pass that automatically fuses operations that can be run efficiently as a single Intel® MKL operation.

Other Optimizations

Intel has also tweaked a number of TensorFlow framework components to enable the highest CPU performance for various deep learning models. For example, Intel developed a custom pool allocator using existing pool allocator in TensorFlow to ensure that both TensorFlow and Intel® MKL share the same memory pools (using the Intel® MKL imalloc functionality) and Intel doesn't return memory prematurely to the operating system, thus avoiding costly page misses and page clears. In addition, Intel carefully tuned multiple threading libraries (pthreads used by TensorFlow and OpenMP used by Intel® MKL) to coexist and not to compete against each other for CPU resources, thus improving the comprehensive utilization rate of resources.

To find out more, visit:

- https://www.intel.ai/tensorflow/?_ga=2.231295069.330745958.1563951842597697079.1551333838&elq_cid=4287274&erpm_id=7282583
- https://www.intel.ai/improving-tensorflow-inference-performance-on-intel-xeon-processors/#gs.v0kayg



Latency Performance of TensorFlow Based on Intel® oneDNN[1]

[1] System Configuration: Intel® Xeon® Platinum 8180 Processor @2.50 GHz; OS: CentOS Linux 7 (Core); TensorFlow Source Code: https://github.com/tensorflow/tensorflow; TensorFlow Commit ID: 355cc566efd2d86fe71fa9d755ceabe546d577a

Intel® Distribution for Python

Intel® Distribution for Python is a powerful software development tool kit developed by Intel. It provides everything needed to write Python native extensions, including C and Fortran compilers, math libraries and analyzers, and integrates multiple high-performance data analysis and math libraries, such as NumPy, SciPy, Scikit-learn, Pandas, Jupyter, matplotlib, mpi4py.

Intel® Distribution for Python is one of the important toolsets of Intel® Parallel Studio XE, with many features and high efficiencies:

- Accelerate compute-intensive applications including numbers, science, data analysis, machine learning, etc. by providing out-of-the-box tools such as uMath, NumPy, SciPy and Scikit-learn.
- Integrated with Intel® performance library (such as Intel® MKL); built-in latest vectorization instructions, such as Intel® AVX-512 and multithreading instructions, Numba and Cython; access composable parallelism with multithreading building blocks library Intel® TBB; unlock Python's parallel application function based on multi-core processor, thus improving the performance of Python program running on platforms based on Intel® architecture, and ensuring good system compatibility without any code changes.
- Support Python2.7, Python3.6 and the latest generation of Intel® processors. It provides optimized deep learning libraries and machine learning libraries such as TensorFlow, Caffe, like Support Vector Machines (SVM) and K-means prediction, random forests and XGBoost algorithms, so as to construct and expand production ready algorithms for workloads such as scientific computing and machine learning.

The benchmark tests compare the efficiency of the Intel® Distribution for Python with the Scikit-learn toolkit (a toolkit widely used for numerical calculation, scientific calculation and machine learning) in other open source Python, and show that the performance indicators of Intel® Distribution for Python have been significantly improved (see the figure below. The higher the efficiency, the faster the function and the closer to native C speed). For example, the efficiency of algorithms such as K-means clustering and linear regression in Intel® Distribution for Python can reach 90% of the efficiency of C in Intel® DAAL.

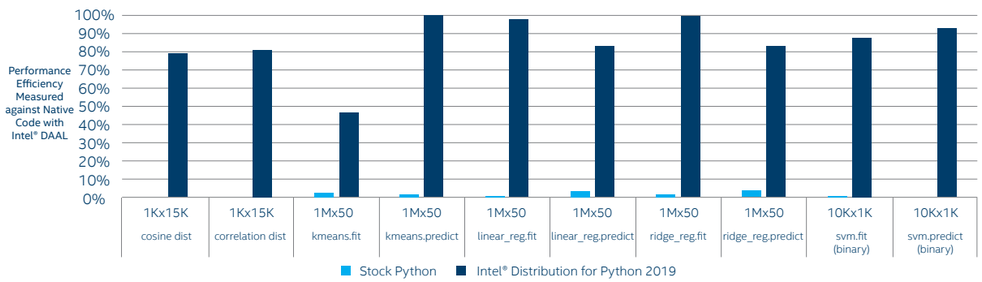
Simple deployment and ease of use are also one of the major features of Intel® Distribution for Python - by using Conda Package Manager and Anaconda Cloud, users can install the core Python environment with only one command:

- **conda install intelpython3 -c intel**

In addition, the Intel® Distribution for Python is a pre-built binary file and can also be obtained through various channels such as pip, Docker images, YUM and APT repos.

To find out more, visit:

https://software.intel.com/en-us/distribution-for-python



Intel's Optimizations Improve Python Scikit-learn Efficiency Closer to Native Code Speeds on Intel® Xeon® Processors

Intel® Distribution for PyTorch

PyTorch is an open source deep learning library designed to facilitate data scientists and developers improve the efficiency of deep learning training and inference, with a high degree of flexibility, ease of use and high speed of training and inference, and is also very popular in the industry.

PyTorch offers a number of excellent features based on its simple, flexible architecture. For example, it provides automatic derivation, which makes model building simple and fast; it uses dynamic models, which can be tuned at any part of the execution process, greatly improving ease of use; it can automatically calculate gradients and update network parameters; it provides flexible training methods, allowing users to customize the communication methods to implement new communication algorithms; it provides a hybrid Python and C++ front-end, which can be used for training in Python front-end and deployment in C++ front-end. In addition, PyTorch is supported by a strong community and rich resources.

To achieve high performance and efficiency on CPUs, Intel introduced the Intel® Distribution for PyTorch. This optimized version leverages Intel® MKL, which enables it to take full advantage of the CPU's parallel computing power and vectorization techniques to optimize matrix multiplication performance, and Intel® MKL-DNN, which enables it to maximize the performance of computing commonly used in deep learning, such as convolution and normalization by utilizing CPU's parallel computing power and vectorization techniques and the efficient use of the CPU's on-chip cache. For the computational graph characteristics at hierarchical or node-level of deep learning networks, Intel optimizes various layers at the node level, such as convolution, matrix multiplication, ReLU, and pool, to minimize data transfer and ensure efficient use of SIMD instructions, execution units, registers, and memory cache hierarchies; at the computational graph level, Intel uses various data ordering policies and layer fusion to optimize groups of nodes, such as fusing ReLU to convolution so that data still performs ReLU operations at the last convolution cycle in the registers. In addition, Intel® MKL-DNN is integrated into the PyTorch back-end by leveraging the key DNN layer of Intel® MKL-DNN API.

With multiple in-depth optimizations, when running on the Intel® Xeon® Platinum 8280 processor, Intel® Distribution for PyTorch has been tested to deliver 7.7X, 47X and 23.6X performance improvements on ResNet50, Faster R-CNN and RetinaNet, respectively.[1]

In addition, when using Intel® Distribution for PyTorch, users are often not required to modify the original Pythorch scripts and codes.

To find out more, visit:

https://software.intel.com/content/www/us/en/develop/articles/intel-and-facebook-collaborate-to-boost-pytorch-cpu-performance.html?wapkw=Pytorch&elq_cid=4287274&erpm_id=7282583

[1] https://software.intel.com/content/www/us/en/develop/articles/intel-and-facebook-collaborate-to-boost-pytorch-cpu-performance.html?wapkw=Pytorch&elq_cid=4287274&erpm_id=7282583

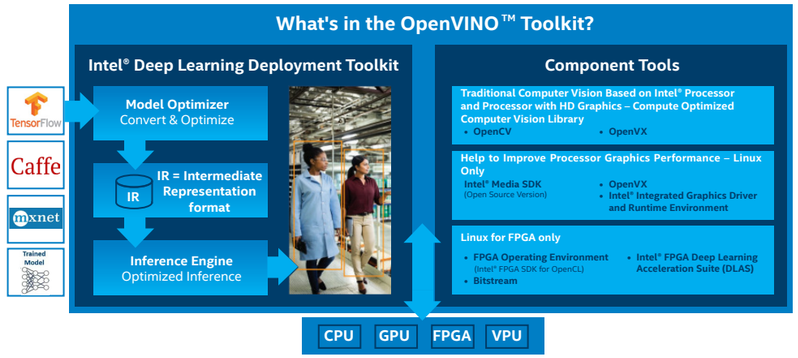
OpenVINO™ Toolkit

OpenVINO™ Toolkit is a software toolkit introduced by Intel to accelerate deep learning inference and deployment, which is used to accelerate high-performance computer vision processing and application. The tool allows for heterogeneous execution, supports Windows and Linux systems, and Python/C++. It can effectively promote the in-depth application of computer vision technology in fields from intelligent cameras, video surveillance, robots, to intelligent transportation, intelligent healthcare, and so forth.

This toolkit is designed to increase performance and reduce development time for computer vision solutions. It simplifies access to the benefits from the rich set of hardware options available from Intel which can increase performance, reduce power, and maximize hardware utilization – letting you do more with less and opening new design possibilities.

By extending workloads across Intel® hardware (including accelerators) based on deep Convolutional Neural Network (CNN), the OpenVINO™ toolkit can rely on chips such as Integrated GPU with Intel® processors, FPGA and Intel® Movidius™ VPU to enhance the features and performance of the vision system. The newly released version of OpenVINO™ has been able to support the 2nd Generation Intel® Xeon® Scalable Processor, and improve inference performance by using Intel® AVX-512 and Intel® DL Boost with VNNI. It can help customers to quickly complete hardware product upgrade and algorithm migration without changing software, thus helping them to accelerate the development of high-performance computer vision and deep learning applications on the edge:

- Increase deep learning performance related to computer vision up to 19X on Intel® architecture platform[1];
- Release the performance bottleneck of CNN-based network in edge devices;
- Accelerate and optimize the traditional API implementation of OpenCV and OpenXV visual libraries;
- Run on devices including CPU, GPU, FPGA based on common API interface;
- The OpenVINO™ toolkit optimized for Intel® platform mainly includes two parts: deep learning deployment toolkit and traditional computer vision toolkit. The deep learning deployment toolkit includes two core components: Model Optimizer and Inference Engine;
- Model Optimizer: converts a given model into a standard Intermediate Representation (IR), and optimizes the model. Support for deep learning frameworks including ONNX, TensorFlow, Caffe, MXNet, Kaldi;
- Inference Engine: supports accelerated operation of deep learning model at hardware instruction set level, and supported hardware devices include: CPU, GPU, FPGA, VPU.



Provide a cross-platform tool to support computer vision and deep learning inference acceleration

Meanwhile, the traditional OpenCV image processing library has also been optimized with instruction set, achieving significant improvement in performance and speed. Computer vision tools include:

- **OpenCV:** Precompiled OpenCV and the new Intel® Photography Vision Library, with features such as face detection/recognition, blink detection, smile detection;
- **OpenVX:** Graphics-based OpenVX implementation that supports traditional CV operations and CNN primitives. Supports Khronos OpenVX neural network extension 1.2;
- **Miscellaneous:** Include OpenCL™ driver and runtime libraries, and media drivers to simplify computer vision SDK working with Intel® Media SDK and Intel® SDK OpenCL™ applications. Intel® MKL- DNN and CLDNN are included and do not need to be downloaded separately.

In addition, as a toolkit designed to quickly and effectively implement computer vision and deep learning in multiple applications, the OpenVINO™ toolkit optimized for Intel® platform currently provides MO files of pre-converted Caffe, TensorFlow and MXNet models, such as VGG-16, VGG-19, Squeezenet, ResNet, Inception, CaffeNet, SSD, Faster-RCNN, and FCN8, and has more than 100 pre-trained models. Software developers and data scientists can use these tools to quickly build personalized deep learning applications, and can use the basic libraries of OpenCV and OpenVX to create specific algorithms and develop customized and innovative applications.

The OpenVINO™ toolkit has made vision a reality on Intel platforms and has helped many users to easily develop and rapidly deploy computer vision applications, demonstrating the great potential of AI solutions in a variety of deep learning application scenarios.

To find out more, visit:

https://software.intel.com/zh-cn/openvino-toolkit

[1] https://software.intel.com/en-us/articles/a-guide-for-setting-up-docker-based-openvino-development-environment-with-ubuntu-system

Intel® Software Guard Extensions (Intel® SGX)

Intel® Software Guard Extensions (Intel® SGX) is a new set of instruction set extension and access control mechanism that support Intel® Xeon® processors E3-1500 v5 and v6, Intel® Xeon® processor E2100 series, Intel® Core™ processor family from 6th generation onwards, and Intel® Celeron® processors J4105 or J4005 (BIOS model with Intel® SGX support). It is designed to effectively prevent application code and data breach or tampering through hardware-based isolation and memory encryption. With Intel® SGX, developers can place applications and sensitive data in "enclaves" within specific hardware to achieve hardware-assisted confidentiality and integrity protection and effectively block access from processes with higher privilege, thereby enhancing security levels and providing several features:

- Enhanced confidentiality and integrity protection
- Remote authentication & supply
- Low learning curve
- Significantly reducing the attack surface

Pushing application security limits

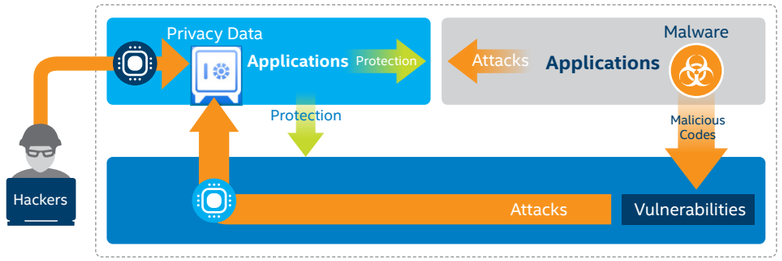
For a long time, developers have been limited by the security features developed and configured by the platform provider for applications. The Intel® SGX now uses a new model that leverages platform and Operating System (OS) strengths to improve security, and allows developers to understand and decide which applications and confidential data need additional protection, thereby unlocking the new performance benefits of chip-level security technologies and effectively facilitating applications protect themselves based on developer settings.

Providing a new approach to security

To help address many security vulnerabilities and system threats, Intel has developed a hardware-based Trusted Execution Environment (TEE) to reduce the attack surface. Intel® SGX provides new Intel® architecture instructions that allow applications to use them to partition areas for greater privacy, as well as to select code and data to prevent direct attacks on code or data stored in memory while it is executing.

Enhancing the effectiveness of federated learning

During federated learning performed on Intel® SGX, AI models and training data are deployed in protected hardware enclaves, which dramatically reduces communication and computing costs associated with application and data encryption and decryption, making execution more efficient. In addition, applications with Intel® SGX support can be developed, integrated and executed on a specific Intel® architecture processor platform, requiring only the installation of drivers and SDK adaptation, with no additional hardware or software environment, no changes to programming, and a lower learning curve. Meanwhile, the TEE hardware solutions based on Intel® SGX run more efficiently than federated learning implementations that adopt techniques such as secure multi-party computation, encryption, and so on.



Block insider attacks on applications to better protect code and data

Intel® SGX Application

The Intel® SGX application contains two parts, an untrusted startup component and a trusted component, with production code running in the "enclave" of the trusted component. Numerous solutions benefit from the additional protection provided by Intel® SGX, including Artificial Intelligence (AI) and Machine Learning (ML) processing, key management, proprietary algorithms, biometric protection, and more. Developers can support a distributed architecture by creating 1 to n enclaves that run collaboratively. While the program is running, Intel® SGX instructions create "enclaves" and execute them to specific areas of encrypted memory, and developers can restrict access to/from this area in order to prevent data breach.

Enclave authentication and data sealing

Intel® SGX supports local authentication or relying party remote authentication between enclaves to ensure that applications are not corrupted. One part of the application is loaded into an "enclave" for code and data measurement. The enclave report is then sent to the remote application owner's server, which in turn verifies that whether the enclave report is generated by a real Intel processor or not.

Data center authentication

Intel® SGX Data Center Authentication Source Code (Intel® SGX DCAP) allows enterprises, data centers and cloud service providers to build and deliver their own authentication services without the need for remote authentication from a third-party provider.

Enabling security model innovations

The fundamental function of Intel® SGX is to provide a higher level of isolation and authentication for the platform's Operating System (OS), application and hardware code, data, and core Internet Protocol (IP) to significantly reduce the likelihood of software attacks, and has been used to enhance the security of a wide range of use cases and applications, including:

- Key Management
- Blockchain
- Privacy-Enhancing Analytics & Workloads
- Applications at Runtime
- Hardware-Enhanced Content Protection
- Enhanced Application & Data Protection
- Edge Computing
- Digital wallet
- Communications & Messaging

Intel® SGX Resources

Intel® SGX Commercial License Information:
https://software.intel.com/sgx/request-license

Intel® SGX Software Development Kit (SDK):
https://software.intel.com/sgx/sdk

Download the Intel® SGX Documentation and Software Development Kit (SDK) at:
https://software.intel.com/sgx

Glossary of Terms Used in This Guidebook

Full Name	Abbreviations
Adaptive Boosting	AdaBoost
Automatically Tuned Linear Algebra Software	ATLAS
Basic Linear Algebra Subroutine	BLAS
Batch Size	
Brain Tumor Segmentation	BraTS
Cardiac Magnetic Resonance	CMR
Classification And Regression Tree	CART
Column Subsampling	
Computed Tomography	CT
Concat Ops	
Constant folding	
Convolution Ops	
Convolutional Layer	
Convolutional Neural Network	
Cosine Similarity	
Cross Validation	
Curse of Dimensionality	
Cyclic Institutional Incremental Learning	CIIL
Deep Supervision	
Deep Learning	DL
Double Data Rate	DDR
Dynamic Random-Access Memory	DRAM
Euclidean Distance	
Feature Extraction	
Feature Map	
Federated Learning	FL
Federated Transfer Learning	
Fully Connected Layer	
Fully Convolutional Network	FCN
Fused Multiply Add	FMA
General Matrix Multiply	GEMM
Geometric Pattern	
GNU Compiler Collection	GCC

Full Name	Abbreviations
Gradient Boosting Decision Tree	GBDT
High Content Screening	HCS
Horizontal Federated Learning	
ImageHub	
Inference Engine	
Institutional Incremental Learning	IIL
Intel® Advanced Vector Extensions	Intel® AVX
Intel® Deep Learning Boost	Intel® DL Boost
Intel® Deep Learning Deployment Toolkit	Intel® DLDT
Intel® Math Kernel Library for Deep Neural Networks	Intel MKL-DNN
Intel® Streaming SIMD Extensions	Intel® SSE
Intel® Ultra Path Interconnect	Intel® UPI
Intermediate Representation	IR
Last Level Cache	LLC
Layer Fusion	
Layer-wise Relevance Propagation	LRP
Learning Rate	LR
Least Absolute Shrinkage and Selection Operator	LASSO
Liquid-Based Cytologic Preparation	LBP
Logistics Regression	LR
Magnetic Resonance Imaging	MRI
Mahalanobis Distance	
Matrix Multiplication	
Model Optimizer	
Multi-Grained Scanning	
Multi-Scale Convolutional Neural Networks	M-CNN
Multi-Scale Prediction	
Non- Uniform Memory Access Architecture	NUMA
Object Detection	
Open Message Passing Interface	OpenMPI
Open Multi-Processing	OpenMP
Open Neural Network Exchange	ONNX
Operations Per Second	OPS

Full Name	Abbreviations
Picture Archiving and Communication Systems	PACS
Pixel Intensity	
Platform as a Service	PaaS
Pooling Layer	
Position-Sensitive RoI Pooling	
Position-Sensitive Score Map	
Positron Emission Tomography CT	PET-CT
Primitive	
Principal Component Analysis	PCA
Proximal Gradient Descent	PGD
Radiology Information System	RIS
Random Forest	RF
Receiver Operating Characteristic Curve	ROC
Region of Interest	ROI
Region Proposal Network	RPN
Reorder Ops	
Resample Ops	
Residual Net	ResNet
Root Mean Squared Error	RMSE
Scanning Electron Microscope	SEM
Single Instruction Multiple Data	SIMD
Skip Connection	
Sliding Window Algorithm	
Software as a Service	SaaS
Standard Uptake Value	SUV
Standardized Euclidean Distance	
Support Vector Machine	SVM
The Thrombolysis in Myocardial Infarction	TIMI
Total Cost of Ownership	TCO
Trusted Execution Environment	TEE
Vertical Federated Learning	
Volumes Of Interest	VOI

Disclaimers:

Software and workload 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 the aforementioned factors may cause test 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 product. For more information, go to www.intel.com/benchmarks.

Tests document performance of components on a particular test, in specific systems. Differences in hardware, software, or configuration will affect actual performance. Consult other sources of information to evaluate performance as you consider your purchase. For more complete information about performance and benchmark results, visit intel.com/benchmarks.

Intel technologies' features and benefits depend on system configuration and may require enabled hardware, software, or service activation. Performance varies depending on system configuration. No component or product can be absolutely secure. Check with your system manufacturer or retailer, or learn more at intel.com.

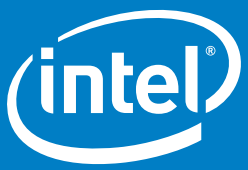
Optimization Notice: Intel's compilers may or may not optimize to the same degree for non-Intel microprocessors for optimizations that are not unique to Intel microprocessors. These optimizations cover SSE2, SSE3, and SSSE3 instruction sets and other optimizations. Intel does not make any warranties as to the existence, functionality or effectiveness of any optimizations on non-Intel microprocessors. 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

No component or product can be absolutely secure.

Cost reduction scenarios described are intended as examples of how a given Intel- based product, in the specified circumstances and configurations, may affect future costs and provide cost savings. Circumstances will vary. Intel does not guarantee any costs or cost reduction.

Intel does not control or audit third-party data. You should review this content, consult other sources, and confirm whether referenced data is accurate.



To accelerate AI implementation, please visit:

Our official website Intel.cn/ai

Micro Blog @ Intel Business

WeChat Intel Biz