CASE STUDY

2nd Generation Intel® Xeon® Scalable Processors
Intel® Distribution of OpenVINO™ toolkit
Intelligent Video Service

Optimizing the Efficiency of Deep Learning Inference, Creating Greater Intelligent Video Services

iQIYI 爱奇艺
悦 享 品 质

Introduction

It has become the consensus of many online video service providers to provide users with a variety of online video services based on Artificial Intelligence (AI) that "caters to their preferences". As a leading company in domestic online video services in China, iQIYI* has been actively promoting convergence of AI applications and video services and has achieved great results in full-process intelligent video services such as intelligent creation, intelligent production and intelligent broadcasting.

As video AI applications are evolving and playing a more and more important role in video services, they pose new challenges to the iQIYI infrastructure. In response to these challenges, iQIYI has combined AI with cloud computing to build an innovative Jarvis* Deep Learning Cloud Platform to meet the requirements of intelligent video services in terms of business resiliency, unified resource scheduling and support for mainstream deep learning frameworks etc.

The successful operation of the new platform depends to a large extent on the greater computing power and higher deep learning efficiency. To help iQIYI further optimize the performance of its Deep Learning Cloud Platform, Intel not only utilized Intel® Xeon® Scalable processors to provide the platform with greater computing power, but also fine-tuned software and hardware for its deep learning inference capabilities based on the technical features of Intel® Architecture processors. These include system-level optimizations with the Intel® Distribution of OpenVINO™ toolkit (OpenVINO™ toolkit). The optimizations have helped iQIYI significantly improve the efficiency of deep learning inference on AI applications and reduce the Total Cost of Ownership (TCO), enabling greater productivity of AI in intelligent video services.

"The use of AI technology can make video services more efficient and intelligent in terms of creation, production, distribution and broadcasting, and give users a better viewing and interactive experience. Intel® Xeon® Scalable processors and OpenVINO™ toolkit not only allowed our Deep Learning Cloud Platform to gain greater computing power, but also significantly improved the efficiency of deep learning inference."

Zhang Lei
Researcher
iQIYI

Benefits of the iQIYI solution:

- The introduction of OpenVINO™ toolkit helps the iQIYI Jarvis Deep Learning Cloud Platform to effectively improve the efficiency of deep learning inference of its AI applications, enabling the optimized acceleration performance of different applications to be enhanced by up to several times to tens of times[1];
- The computing performance enhancements provided by Intel® Xeon® Scalable processors further promotes the inference efficiency of these AI applications.

Watching videos online is an important part of today's internet lifestyle. By the end of 2018, China's online video subscribers have reached 612 million with a 12.8% utilization rate[2]. As the industry chain continues to mature, video service providers are driven to introduce new technologies and capabilities to improve service efficiency and user experience. As a high-quality video-based entertainment service provider with industry influence, iQIYI is actively introducing more AI capabilities to build a full-process intelligent service system for online video.

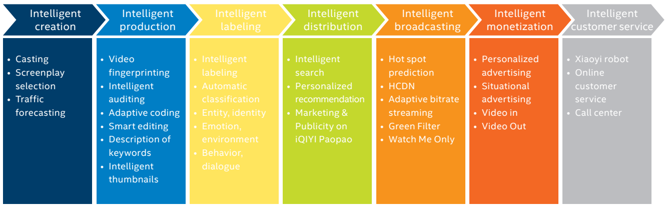

Figure 1. AI applications in iQIYI

As shown in Figure 1, iQIYI uses AI capabilities to implement intelligent transformation around its process of video creation, production, distribution and monetization. Taking "Casting" in video creation as an example, with Natural Language Processing (NLP), iQIYI can extract key information from character and actor information and determine their matching for appropriate roles using the AI algorithm. Another example is that in video broadcasting, the video platform can achieve Adaptive Bitrate Streaming (ABS) with the AI enhanced learning model and improve the viewing experience.

As AI technology plays an increasingly important role in iQIYI video services, it makes more demands on existing infrastructure. Firstly, the explosive growth of AI applications requires the infrastructure to provide rapid and agile deployment capabilities. Secondly, diverse AI models and frameworks require better infrastructure support. In addition, for AI applications deployed in different environments such as the cloud, Content Delivery Network (CDN) and customer premises, how to effectively allocate computing resources to improve their efficiency is also a key concern for iQIYI.

iQIYI has risen to the above challenges by building Jarvis, a cloud-based deep learning platform. Through in-depth technical cooperation with Intel, it has carried out comprehensive software and hardware optimization for Intel® Architecture, ultimately improving the efficiency of AI applications.

Analysis of the iQIYI Deep Learning Cloud Platform

Jarvis Deep Learning Cloud Platform is being used to meet the needs of AI applications. It is divided into four layers from bottom up. As shown in Figure 2, its Intel® architecture-based hardware platform, which consists of high-performance computing, networking, and storage capabilities, forms a solid infrastructure layer for the entire cloud platform. Above the infrastructure layer is a resource management and orchestration layer comprising Apache Mesos* and Docker*, which enables unified management, scheduling and auditing of the underlying resources. At the same time, the containerized operating environment provided by the platform allows users to apply for and release resources on demand, which can effectively improve the resource utilization of the platform.

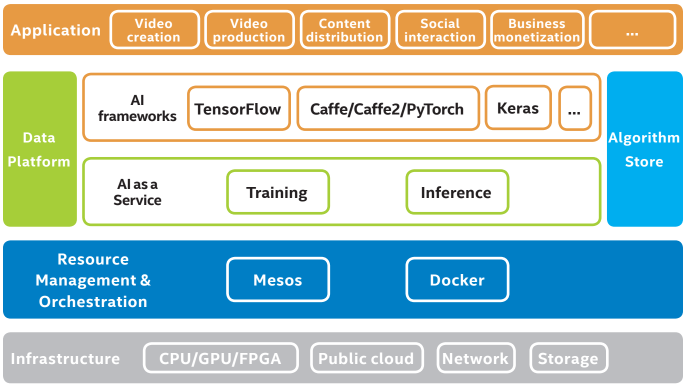

Figure 2. Architecture Diagram for iQIYI Jarvis Deep Learning Cloud Platform

Above the resource management and orchestration layer is the core capability layer of the Jarvis platform. It provides a one-stop AI service with modules such as data platform, algorithm store and AI capabilities (AI training and inferencing). The platform aggregates massive amounts of data from inside and outside iQIYI through methods such as data capture, crowdsourcing collection and importing public datasets. The platform also provides various ways of data labeling, including intelligent labeling and crowdsourcing labeling, for labeling structured/unstructured data such as text, images, video and audio. The algorithm store provides platform users with a variety of AI algorithms, networks and models for displaying and communicating, as well as evaluation of various algorithms' performance so that users can choose which is best for their needs.

For key AI modules, the Jarvis platform supports mainstream deep learning frameworks such as Caffe*, TensorFlow*, Torch*, Keras*, and MXNet* on one hand; while on the other hand, its training module also provides AI applications with a distributed training environment based on the underlying computing resource pool. The inference module provides plentiful AI models and services, allowing users to quickly and easily deploy high-performance inference services. At the top layer of the platform, the application layer, a series of video-based AI applications, such as video creation, content distribution, and commercial monetization, are deployed for internal and external iQIYI subscribers to use.

With the Jarvis Deep Learning Cloud Platform, iQIYI builds a complete AI application deployment process. Initially, data sources from Business Intelligence (BI) platforms, big data analytics, and various AI application requirements will be aggregated into the data system. Users add data to the task for training through the WEB interface, command lines or API interfaces while models acquired or updated through training will be included in the AI model library. Finally, the users choose the appropriate AI algorithm in the algorithm store to implement the inference, and ultimately give feedback to a higher-level application or request by means of HTTP or the like.

Software and hardware optimization based on Intel® architecture

Improving the efficiency of deep learning inference is one of the key capabilities of the iQIYI Jarvis Deep Learning Cloud Platform to enhance the productivity of video services. With the help of Intel, iQIYI fully optimized the deep learning inference capabilities of the cloud platform based on Intel® architecture processors.

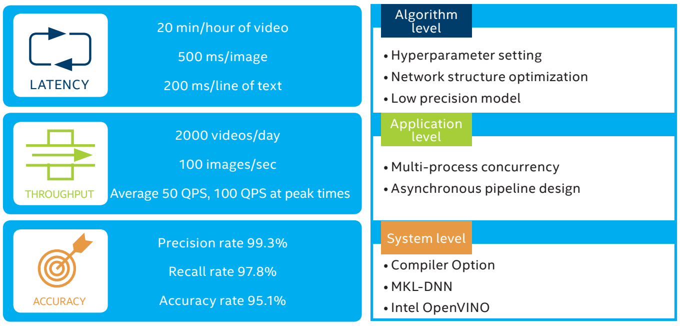

Figure 3. Inference performance optimization indicators and optimization solutions developed by iQIYI

As shown in Figure 3, iQIYI first identifies 3 performance indicators i.e. response latency, throughput and model accuracy, and develops a three-level optimization solution at the levels of system, application and algorithm, among which the algorithm-level optimization focuses on optimizing the deep learning model itself, using methods such as hyperparameter setting, network structure pruning and quantization to reduce the size and computing intensity of the model thereby accelerating the inferencing process. At the application-level optimization, inference efficiency is enhanced by improving the pipeline and concurrency of specific applications and services. In general, deep learning services include not only inferencing, but also data preprocessing, post processing and network request response. Good concurrency design can effectively improve the end-to-end performance of these applications on the server. At the system-level optimization, the introduction of methods such as the Single Instruction Multiple Data (SIMD) instruction set, OpenMP multi-thread library and Intel® MKL/MKL-DNN enables the computing power of the entire platform to be fully accelerated, thereby improving overall efficiency.

At the system-level of optimization, iQIYI has also introduced OpenVINO™, an AI toolkit from Intel for the Jarvis platform.

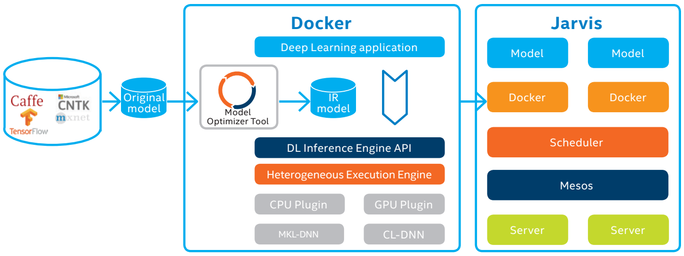

Figure 4. The Jarvis platform inference optimization process based on the OpenVINO™ toolkit

As shown in Figure 4, the OpenVINO™ toolkit first uses a Model Optimizer Tool to transform and optimize the native deep learning model to generate Intermediate Representation (IR), which contains the optimized network topology, model parameters and model variables, and then the Inference Engine reads the IR and performs inference.

As a result of the organic integration of computer vision and AI technology, the OpenVINO™ toolkit accelerates AI inference on different computing resources (including Intel® processors, FPGAs and VPUs) for the Jarvis platform. It includes a deep learning inference acceleration toolkit and a computer vision toolkit that provides excellent support for deep learning frameworks such as TensorFlow, MXNet, and Caffe.

Taking the bullet comments during video broadcasting as an example, the Jarvis platform hides the bullet comments behind the main object in the video through the AI application based on the DeepLab v3 + * Deep Learning Model to prevent those comments from interfering with normal video broadcasting. The DeepLab v3 + Model is a semantic image segmentation model based on a deep convolutional network. It enables the function by applying matting to the image in a single video frame. Compared to traditional computer vision algorithms, this model can adapt to a variety of complex textures and scenarios such that it provides more accurate results and more agile deployment capabilities in scenarios with similar foreground and background colors.

Test data from iQIYI shows that the introduction of the OpenVINO™ toolkit has helped the Jarvis platform increase the inference speed of displaying real-time bullet comments by about five times. Other deep learning models on the iQIYI Jarvis platform have also validated acceleration results that the OpenVINO™ toolkit brings about. As shown in Figure 5, the efficiency of the facial recognition application has also been increased by about four times, and that of pornographic content detection by about six times. And in text detection applications, inference performance is improved by as much as 11 times after optimization[3].

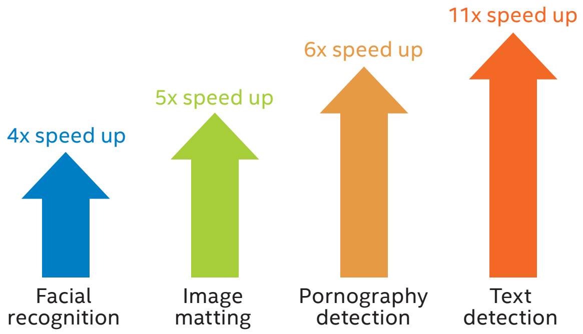

Figure 5. OpenVINO™ toolkit enhances the efficiency of the Jarvis platform

It is worth mentioning that the more powerful Intel® Xeon® Scalable processors can further enhance the performance of OpenVINO™-based AI applications. The new processors tend to integrate more cores, and when the platform performs offline inferencing work, the inference throughput increases linearly with the increasing number of processor cores. On the other hand, the more optimized instruction set in the new processors, such as Intel® AVX-512 built into Intel® Xeon® Scalable processors, also provides greater performance acceleration. A data comparison from iQIYI shows that with the same OpenVINO™ toolkit performance is doubly accelerated using Intel® Xeon® Gold 6148 processor as compared to Intel® Xeon® E5-2650 v4 processor[4]. With the release and deployment of the 2nd Generation Intel® Xeon® Scalable processors on the iQIYI Jarvis Cloud Platform this year, performance will surely move up to the next level.

Conclusion

Various optimization methods and tools represented by the OpenVINO™ toolkit have been applied in more than 10 AI applications on the iQIYI Jarvis Deep Learning Cloud Platform and deployed in thousands of cores[5]. Feedback from front-line users of the iQIYI Jarvis platform indicates that these optimization methods have effectively improved the performance of many video AI applications and significantly increased productivity in video service.

Looking forward, iQIYI will continue to work with Intel to further optimize its deep learning efficiency and plans to add more heterogeneous computing resources to its Jarvis Deep Learning Cloud Platform to accelerate specific tasks. At the same time, both parties plan to make full use of computing resources in terms of service flexibility, scheduling optimization and automatic parameter selection to enable deep learning inference services to acquire more flexible deployment capabilities.

[1,3,4,5] Data cited from Intel Blog (https://software.intel.com/en-us/articles/optimization-practice-of-deep-learning-inference-deployment-on-intel-processors) and the speech presented by iQIYI at the Intel Artificial Intelligence Conference 2018 titled "iQIYI AI Video Cloud Architecture and Optimization Practices". All relevant performance data comes from the internal testing conducted by iQIYI. For more details, please contact iQIYI.

[2] This data is cited from the 43rd "Statistical Report on the Internet Development in China" issued by CNNIC (China Internet Network Information Center).

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

Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors. Performance tests, such as SYSmark and MobileMark, are measured using specific computer systems, components, software, operations and functions. Any change to any of those factors may cause the results to vary. You should consult other information and performance tests to assist you in fully evaluating your contemplated purchases, including the performance of that product when combined with other products. For more complete information visit www.intel.com/benchmarks.

Performance results may not reflect all publicly available security updates. See configuration disclosure for details. No product or component can be absolutely secure.

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 product or component can be absolutely secure. Check with your system manufacturer or retailer or learn more at intel.com.