CASE STUDY

Intel® Optimization for TensorFlow
Intel® Optimization for Caffe
Intel® MKL-DNN
Intel® AVX-512
AIaaS
IaaS

Kingsoft Cloud Upgrades Cloud Service for AI Developers by Bundling Optimized Hardware and Software for Intel® Xeon® Platform

Kingsoft Cloud

"Kingsoft Cloud has been striving to provide first-class cloud services for AI R&D. This not only needs a high-performance hardware platform as a foundation, but also hinges on collaboration and full optimization of software and hardware. By deploying various Intel advanced hardware products and with the introduction of Intel® Optimization for TensorFlow, Intel® Optimization for Caffe and other optimized frameworks, we offer developers enhanced IaaS while significantly reducing their investment on system deployment and optimization. This allows them to focus more on the AI business itself."

Feng Yang
Cloud Computing R&D Director
Kingsoft Cloud

Employing cloud services to accelerate the research and development (R&D) of Artificial Intelligence (AI) applications has become the rule of thumb for many AI development teams. As a global provider of premium cloud services, Kingsoft Cloud endeavors to provide superior Infrastructure as a Service (IaaS) via an array of high-performance cloud servers like Kingsoft Elastic Compute and cloud physical hosts such as Elastic Physical Compute to help users gain a first-mover advantage in voice, image, video and many other AI application scenarios.

To help developers increase their AI R&D efficiency, Kingsoft Cloud, together with its strategic partner Intel, deployed Intel® Xeon® Scalable processors and other advanced hardware products in its cloud instances and introduced optimizations of AI frameworks including Intel® Optimization for TensorFlow and Intel® Optimization for Caffe*. This bundling of optimized hardware and software for AI enhanced Kingsoft Cloud's IaaS capabilities to better support AI workloads. Users now don't have to bother with complicated configuration and fine-tuning of underlying AI frameworks, instead they get optimal performance on the Intel® Xeon® Scalable processor-based cloud infrastructure with this one-stop service.

Tests by Kingsoft Cloud show that the performance of multiple optimized AI frameworks improved by several times, even up to dozens of times in various deep learning models. This shows that Kingsoft Cloud's enhanced IaaS, bundling optimized hardware and software for the Intel® Xeon® Scalable Platform, can provide outstanding performance for AI R&D in different application scenarios and accelerate developers' AI R&D progress.

Advantages of Kingsoft Cloud solutions[1]:

- The multiple prepackaged Intel optimizations for deep learning frameworks enable users who conduct AI R&D with Kingsoft Cloud to reduce time and effort downloading, deploying and optimizing the relevant frameworks to be able to devote more resources to AI work;
- Intel Optimization for TensorFlow enhances the performance of Kingsoft Cloud instances built on Intel® Architecture-based processors in various deep neural network (DNN) models by 2.45-2.89 times[1];
- With the introduction of Intel Optimization for Caffe, the performance of Kingsoft Cloud instances built on Intel Architecture-based processors improved in various DNN models. In ResNet50, performance improved by nearly 30 times[1].

Delivering High-performance IaaS for AI

The ever-evolving public cloud services are playing an increasingly important role in AI R&D. Featuring flexible resource allocation and high scalability, public cloud services enable agile scheduling of computing power, algorithms and data needed for AI R&D and can raise efficiency. As a result, more and more AI development teams are turning to cloud services for their AI R&D and innovation.

To provide users with more efficient and cost-effective IaaS services, Kingsoft Cloud and Intel are working together to introduce Intel Xeon Scalable processors, Intel® Optane™ DC SSDs, 25 GbE Intel® Ethernet network adapters and other cutting-edge hardware products and technologies into Kingsoft Cloud instances (cloud servers, cloud physical hosts, etc.) as the foundation on which high-performance IaaS capabilities can be built.

Take the Intel® Xeon® Platinum 8168 processor deployed by Kingsoft Cloud as an example, it features optimized microarchitecture, up to 24 cores and 48 threads, delivering higher computing power and scalability for computation-intensive AI inference workloads. Meanwhile, its Intel® Advanced Vector Extensions 512 (Intel® AVX-512) can concurrently process 16 single-precision floating-point numbers, doubling the single-precision floating-point number processing capability compared with the previous generation Intel® Advanced Vector Extensions 2 (Intel® AVX2). It does this via more fused multiply-add (FMA) units, offering a significant advantage when it comes to intensive AI vector computation.

In parallel, Intel® processor platforms are reinforcing these advantages through continuous evolution. The arrival of new 2nd generation Intel® Xeon® Scalable processors featuring integrated Intel® Deep Learning Boost technology will improve the performance and scalability of the overall architecture of Kingsoft Cloud instances. Along with other hardware technologies and products, this will lend weight to AI development teams in various AI application scenarios, such as voice, image and video.

Intel® Optimization for AI Frameworks

But does it follow that enhancing hardware performance leads to an equivalent improvement in AI work efficiency? Observations from Kingsoft Cloud indicate that AI development teams needs to ensure optimal efficiency by driving the installation, deployment and fine-tuning of deep learning frameworks after determining their high-performance hardware devices. This is a waste of precious time and implies resource wastage if the effects of attempted optimization become less than desirable. For example, if a user were to apply for a cloud physical host instance with 24 vCPUs (virtual processor cores), only 50% of the processor cores may be able to be fully utilized if there is insufficient parallel processing capability of the native AI frameworks. As a result, the user would need to apply for more cloud instances to meet their needs, lowering the efficiency of the systems and increasing Total Cost of Ownership (TCO).

To resolve this problem, Kingsoft Cloud is working with Intel to offer users Intel optimization for AI in which multiple optimized deep learning frameworks are integrated. This enhanced IaaS cloud service for AI R&D on Intel Xeon Scalable processors frees developers from downloading, installing, configuring and fine-tuning these frameworks and enables higher performance.

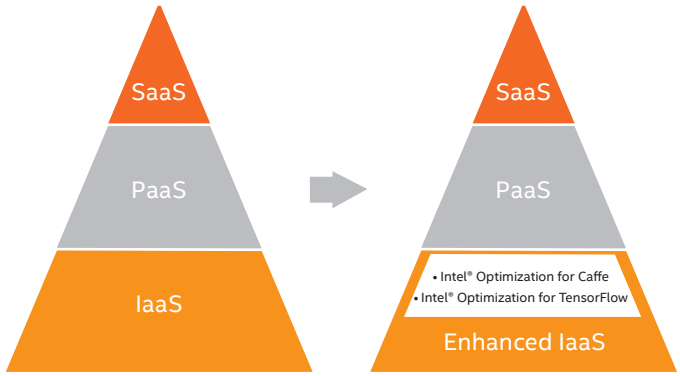

Figure 1. Enhanced IaaS cloud services for AI R&D

Take Intel Optimization for TensorFlow as an example, it provides support for DNN primitive in Intel® Math Kernel Library for Deep Neural Networks (Intel® MKL-DNN). This includes an array of general-purpose AI computing processes such as 2D convolution, inner product/matrix multiplication, batch normalization, ReLU activation and multidimensional transposition. When Intel Optimization for TensorFlow is deployed on Intel Xeon Scalable processor platforms, Intel® MKL-DNN primitive is utilized to help users quickly build-up the necessary functional modules.

Intel Optimization for TensorFlow also uses code refactoring to vectorize massive computing processes (such as convolution and matrix multiplication) which are necessary for deep learning computation, leaving computation to Intel AVX-512 to leverage its strength in vector computation. Apart from this, Intel Optimization for TensorFlow can also schedule idle processor cores to further magnify the multicore power of Intel Xeon Scalable processors.

Similarly, Intel Optimization for Caffe makes full use of Intel MKL-DNN to accelerate various computing processes in AI workloads. For example, highly vectorized and threaded building blocks in Intel MKL-DNN are used to implement convolutional neural network models in C and C++ interfaces and further enhance AI inference performance with converged technology solutions like layer fusion.

Performance Comparisons Before and After Optimization

To verify that the performance of cloud host instances were enhanced with the introduction of Intel optimization for AI, Kingsoft and Intel conducted a series of tests on a Kingsoft Cloud general-purpose N3 instance , focusing on ResNet50 (a residual neural network), ResNeXt50 (an upgraded residual neural network), Inception-V3 (a convolutional neural network), SSD-MobileNet (an object detection network) and Wide & Deep (a classic recommendation algorithm based on MovieLens-1M datasets) network models. These neural network models are widely used for image segmentation, content recommendation and other common AI scenarios.

Initially, the contribution to AI inference performance between Intel Optimization for TensorFlow and native TensorFlow was compared in four DNNs: ResNet50, Inception-V3, SSD-MobileNet and Wide & Deep. The batch size of ResNet50, Inception-V3 and SSD-MobileNet was set at 1, while the batch size of Wide & Deep were set at 256.

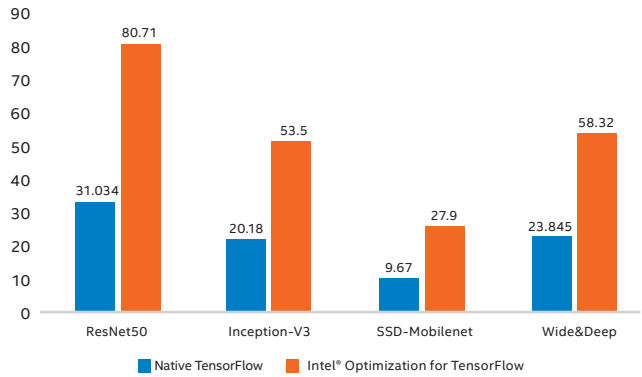

Figure 2. Comparison between native TensorFlow and Intel® Optimization for TensorFlow in their contribution to AI inference performance in different DNNs

The comparison results shown in Figure 2 indicate that Intel Optimization for TensorFlow improved AI inference performance of a cloud instance differently in four DNNs, compared to native TensorFlow. The performance was improved by 2.89 times in SSD-MobileNet.

In another set of tests, Kingsoft and Intel compared the performance of forward propagation implemented by Intel Optimization for Caffe and BVLC Caffe in ResNet50, Inception-V3, SSD MobileNet and ResNeXt50. The batch size in four neural networks was set as 1.

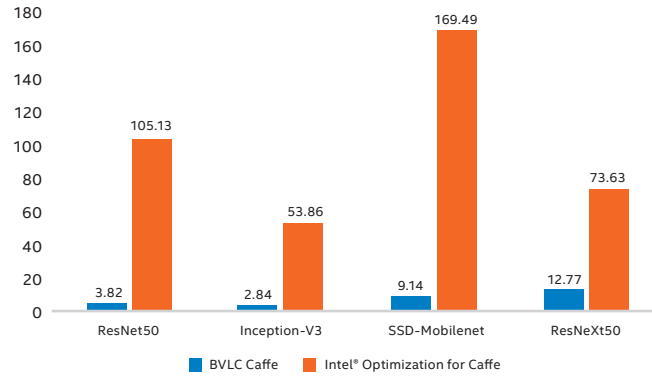

Figure 3. Comparison between native Caffe and Intel® Optimization for Caffe in their contribution to AI forward propagation performance in different DNNs

It can be seen from the results shown in Figure 3 that BVLC Caffe's contribution to forward propagation performance in all four DNNs was below expectation. in contrast, the performance was improved by several times or even dozens of times in these DNNs with the help of Intel Optimization for Caffe. In ResNet50, the improvement reached an incredible 27.5 times.

Conclusion

By offering users both high-performance hardware infrastructure and AI software (comprising multiple optimized deep learning frameworks) in a bundle, Kingsoft's enhanced IaaS provides AI development teams with a one-stop solution featuring higher performance, more comprehensive technological solutions and better scalability. This allows them to allocate more of their resources to the application R&D and to the business itself meaning that they can both increase development efficiency and lower TCO.

Kingsoft Cloud and Intel will continue to extend and enhance their technological cooperation around how cloud services can improve AI R&D efficiency in the future. With the deployment of 2nd generation Intel Xeon Scalable processors, Intel® Optane™ DC persistent memory and other new hardware products in Kingsoft Cloud, both parties will focus on how to harness the value of Intel® Deep Learning Boost and high-density memory cloud instances for AI to enable more research, exploration and productization.

Tips

Intel® Optimization for TensorFlow: TensorFlow is a deep learning open source framework widely used in AI-related areas. It provides support for workloads including computer vision, speech recognition and natural language processing (NLP). To improve the operating performance of native TensorFlow on Intel® Architecture-based processor platform, Intel worked with its partners to apply significant optimizations. These included utilizing the Intel® AVX-512 instruction set more effectively, increasing processer core utilization to achieve higher performance, implementing parallelization on the designated layer or function, or between layers, balancing the use of prefetch module and cache module technologies and improving the data format of spatial and temporal localities. With these optimizations in place, Intel® Optimization for TensorFlow sees impressive enhancement in performance compared with native TensorFlow.

Intel® Optimization for Caffe: Caffe is a deep learning framework developed by Berkeley Vision and Learning Center (BVLC) and community contributors and comes with a large quantity of pretrained models. Besides providing powerful vision, speech and multimedia support for AI applications, it also supports using OpenCV (a widely adopted computer vision library) to boost the computer vision function of mobile devices. Intel Optimization for Caffe inherits all the merits of BVLC Caffe, while providing functionality and multinode distributed training and scoring for Intel optimization. With code vectorization, it can utilize processor resources efficiently, improving function call performance, lowering the complexity of algorithms and cutting the amount of computations. At the same time, this version introduces code optimizations for processors and systems as well as OpenMP code parallelization technology which, together, significantly improve its performance compared with BVLC Caffe.

For more information, please refer to these URLs:

https://github.com/IntelAI/models

https://github.com/intel/caffe

https://www.intel.ai/tensorflow-optimizations-intel-xeon-scalable-processor/

https://software.intel.com/en-us/articles/intel-optimization-for-tensorflow-installation-guide

https://software.intel.com/en-us/mkl/documentation/view-all

https://www.intel.com/content/www/us/en/architecture-and-technology/avx-512-overview.html

[1] The data cited to demonstrate the advantages of the solutions are extracted from the section “Performance Comparisons Before and After Optimization” herein on tests. Read the section for more details.

[2] Intel® Xeon® Platinum 8168 Processor adopted by Kingsoft Cloud’s general-purpose N3 instance has a CPU frequency of 2.70 GHz and uses DDR4 DRAM. The version of Intel® Optimization for TensorFlow used in the tests is r1.12. The version of Intel® Optimization for Caffe used in the tests is 1.15. The data used in the tests here are captured from the cloud physical host instance with 24 vCPUs. For more details, please refer to https://marketplace.ksyun.com/products/10291