Case Study

intel®

TensorFlow Optimization by Meituan Based on Intel® Xeon® Scalable Processors

In an increasingly competitive internet market, leveraging artificial intelligence (AI) to drive business innovation and provide users with more accurate and personalized internet services has become a preferred approach to gain competitive edge. As China's leading e-commerce platform for lifestyle services, Meituan has established an experienced AI team, providing robust AI capabilities to its entire set of operations such as business site selection, traffic driving, delivery services, operational management, supply chain finance, and marketing.

Nevertheless, with a fast-growing number of users, ever-evolving intelligent businesses, as well as increasing scale and complexity of AI models, its business system faces mounting performance challenges. In response, Meituan has re-architected its infrastructure and optimized its software. Consider applications using TensorFlow, an open source deep learning framework, for instance. Meituan has conducted in-depth optimization of the support for large-scale sparse parameters, training mode, distributed communication, pipeline, and operator fusion on Intel Xeon Scalable processors. It has also adopted the recommended optimization solution from Intel. As a result, distributed scalability has been boosted over tenfold in its recommendation system scenarios.[1]

Challenge: Performance Bottleneck of TensorFlow in Large-Scale Applications

Thanks to the exponential growth of data, emerging machine/deep learning algorithms and greater computing power, AI has entered a phase of aggressive development, with innovations and new application implementations rising sharply. For the internet industry, AI will subvert the way resources and businesses are run. Against such a backdrop, internet giants have stepped up their efforts in AI and doubled down their investment in deep learning model training and inferencing.

As the second generation of AI-powered learning system developed by Google, TensorFlow can handle many deep learning algorithm models. Known for its performance, open-source features and high scalability, it has already become an indispensable tool in deep learning research and applications.

TensorFlow has released Wide & Deep Learning, a deep learning algorithm model for recommendation scenarios. The model is designed for general-purpose large-scale regression and classification with sparse inputs, for example, recommendation systems, search and ranking.

To further empower applications like the recommendation system with AI, Meituan used TensorFlow for model training and a distributed computing approach for model computing and parameter upgrading with massive parameters. However, with the evolution of Meituan's businesses, the model for its recommendation system is growing rapidly both in scale and complexity. Samples to be trained have increased from tens of billions to hundreds of billions, and the number of sparse parameters from several hundred to several thousand, both registering a nearly tenfold growth, while the total number of parameters has grown from several hundred million to several dozen billion, up by 100 to 200%.[2] Meanwhile, Meituan's TensorFlow-based model is becoming more and more complex, resulting in an over ten times increase in the single-step calculation time.[3]

In large-scale applications, the official version of TensorFlow is plagued by a series of issues, including waste of memory resources due to the representation of parameters by Variables, poor scalability for thousands of nodes, and the inability to support online deep learning training with large-scale sparse parameters. These have led to a serious performance bottleneck, which not only causes soaring total cost of ownership (TCO), but also becomes a drag for upper-level businesses.

Scaling-up infrastructure is an obvious solution to break through this bottleneck; however, it will also add pressure to TCO and exacerbate the overall complexity of the system. Another solution is to conduct optimizations at the system and software level, which is more cost-effective and feasible in this case. After analyzing the TensorFlow framework and its own business positioning, Meituan discovered that the load balancing of distributed clusters, its incumbent communication mechanism, latency, and single-instance performance are all the key areas to be optimized.

[1] Data cited from Meituan's internal testing.

[2,3] Data provided by Meituan.

Solution: TensorFlow Optimization by Meituan Based on Intel® Architecture

Currently, Meituan's TensorFlow system is mainly built on Intel® Xeon® Scalable processor-based server clusters and uses CPU for TensorFlow model training. As a workload-optimized platform with built-in AI acceleration, Intel Xeon Scalable processors deliver world-class performance and memory bandwidth for high performance computing workloads, AI applications, and high-density infrastructure. Moreover, its built-in Intel® Deep Learning Boost (Intel® DL Boost) with Vector Neural Network Instructions (VNNI) brings enhanced AI inferencing performance, making it an optimal choice of infrastructure to run deep learning applications.

In addition to the outstanding performance, CPU-based servers are more flexible and agile for elastic scaling of businesses, and their easy-to-deploy and easy-to-manage features are ideal for dynamic resource demands in various business scenarios. With the Intel Xeon Scalable processor as the underlying infrastructure, Meituan has applied the asynchronous training mode of the TensorFlow Parameter Server (PS) to support its business requirements for distributed training in the recommendation system.

For higher performance, Meituan has conducted multiple optimizations including single-instance performance optimization and distributed computing optimization .

- **Throughput Optimization for Unit Computing Power**

In Meituan's TensorFlow system, each node undertakes a massive computation load, so, Meituan plans to further unleash the potential of Intel Xeon Scalable processors to boost system performance with limited computing resources. To that end, Meituan identified some high-frequency operators through its CAT (Central Application Tracking) system and had them analyzed with Intel® VTune™ Profiler, a visualized performance analysis tool. It then made targeted optimizations based on the results of analysis. Engineers from Intel helped evaluate the performance of these operators on the next-generation Intel Xeon Scalable processors, and optimize the select operators like matmul, Pad and Unique using Intel® Advanced Vector Extensions 512 (Intel® AVX-512) and parallelization technology. Take Unique & Dynamic Partition operator fusion as an example. In the TensorFlow PS architecture, all the shared parameters including Embedding vectors are stored on the PS and interact with the Worker through network. For the Embedding query, which usually takes place in a large-scale sparse scenario, the Unique operation, followed by DynamicPartition, is performed before the query. Usually, these two operations will be conducted with the existing TensorFlow operators. However, as the memory allocation policy used by the built-in Unique operator is inefficient, the HashTable created would be overly large and very sparse. Compounded by the redundant data traversal with Unique and Dynamic Partition operators, the operations would be prohibitively time-consuming. To address the problem, Intel engineers helped Meituan evaluate the performance of multiple HashTable implementations and provided a parallelization solution. By working with Intel engineers, Meituan chose the Robinhood HashTable to replace the implementation in TensorFlow, simplifying the logic implementation by combining the operators around Unique and Partition sessions of Embedding ID. With these optimizations, the Unique single operator has gained a 51% increase in speed while the end-to-end performance of the real-world models can be improved by around 10%, and the total number of operators is reduced by 4%.[4]

Within its TensorFlow system, Meituan has also applied Intel® oneAPI Math Kernel Library for Deep Neural Networks (Intel® oneMKL-DNN), which uses the SIMD (Single Instructions/Multiple Data) instructions through vectorization and allows for efficient use of multiple cores across multiple threads. This enables maximum cache and computing power utilization in modern CPUs and boosts the effectiveness of instruction sets, usually resulting in better computing performance in deep learning workloads. To leverage the Intel architecture fully and maximize performance, the TensorFlow library has already been optimized with Intel oneMKL-DNN primitives.

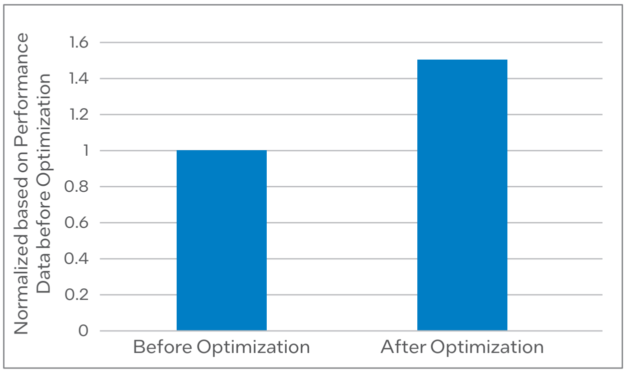

Figure. Performance before and after Unique single operator optimization (Higher is Better)

[4] Data cited from Meituan's internal testing.

Optimization for Distributed Computing

After studying its TensorFlow-based recommendation system, Meituan found that the time for every single-step training would increase instead when the Parameter Servers have scaled to a certain level. The core reason is that Worker single-step training must be accomplished in sync with all PS communications, so N communication links will be added with every additional PS, resulting in significant latency. With millions of or even tens of millions of steps to be accomplished in every training, the latency for communication links overrides the benefits from concurrent computing power of PS's.

To address the issue, the key is to optimize distributed computing with limited PS instances. To that end, Meituan optimized distributed load balancing, distributed cluster communication mechanism and latency.

Distributed load balancing optimization: The Adam optimizer in native TensorFlow could cause imbalanced PS load, which means the load of requests on a certain PS might be much higher than that on others. To solve the problem, Meituan created a β parameter redundantly for the Adam optimizer on each PS, and calculated t and alpha values locally to remove hot spots caused by uneven loads. On an internal business model of Meituan, removing the β hot spot has brought a performance boost by about 9%. Moreover, as global dependence on β is removed, this optimization also improves the scalability of the PS architecture, delivering more efficient acceleration when expanding the number of Workers.

Communication Optimization: Meituan optimized the communication mechanism based on RDMA, including optimization for Memory Registration, the introduction of an RDMA static allocator, and load balancing between Multi RequestBuffer and CQ, which have led to 20% to 40% higher performance in multiple training scenarios.[5]Compared with the communication layer implementation modified by TensorFlow Seastar, the above-mentioned optimization has registered 10% to 60% performance improvement on several of Meituan's business models.

Latency Optimization: First, Meituan aggregated sparse domain parameters, including Embedding, m, v, and the low-frequency filtering counters, as the Value of Hashtable. This significantly reduces the operation frequency of sparse parameters, alleviating the pressure for the PS. After that, Meituan optimized the Embedding pipeline by creating a controllable EG/MG concurrent pipeline training mode and made it transparent to users, who can enable the Embedding pipeline function with just a line of code. By now, the Embedding pipeline has delivered a 20% to 60% performance gain under the CPU PS architecture in one of Meituan's business trainings.[6]

Benefits: Improving Performance and Cost-effectiveness for the TensorFlow-based Recommendation System

Generally, optimization for TensorFlow based on Intel architecture has enabled Meituan to support large-scale sparsity on the TensorFlow architecture. The in-depth optimization performed in several aspects has enabled highly efficient distributed training with hundreds of billions of parameters and samples, unleashed the maximum potential of CPU in deep learning training and improved the performance and cost-effectiveness of the TensorFlow-based Recommendation System.

Specifically, this optimization practice has achieved the following results:

- The new system now supports near linear acceleration for models with hundreds of billions of parameters and the distributed training of thousands of Workers, enabling the full-year sample data to be trained within a single day as well as online deep learning capabilities.
- It now features more friendly architectures and interfaces, and is widely used in Meituan's businesses including food delivery, community group buying, search, ad platform, and Dianping Feeds, etc.

The optimization proves that the Intel Xeon Scalable processor is an optimal platform for deep learning training and inferencing. With optimizations for unit computing power throughput and distributed computing, it has enabled significant performance increase in TensorFlow model training without huge investment in hardware.

[5,6] Data cited from Meituan's internal testing.

Looking Ahead: AI Fueling Faster Digital Innovation in Businesses

Intel provides a wide portfolio of hardware and software products for the fast-growing AI applications. Intel Xeon Scalable processors allow users to do data pre-processing and analytics and run AI applications on the platform or infrastructure with a single architecture. With a set of oneAPI-optimized software tools like the OpenVINO™ toolkit, users can enjoy the "write once, deploy anywhere" flexibility, which helps lower the barrier of AI development, and improve the time-to-market of AI applications, while fully leveraging the existing server resources to save costs.

In the face of the AI surge, Intel will continue its collaboration with partners, such as Meituan, by offering its cutting-edge capabilities in computing, storage, network and others to empower business innovation and speed up the development of the AI industry and adoption of AI technology. It will also provide end users with more accurate and personalized services, lower AI implementation costs and technological barriers while improving performance, all to enable industries to go smart.

About Meituan

The mission of Meituan is "We help people eat better, live better". Meituan focuses on the "Retail + Technology" strategy and joins hands with merchants and partners to enable customers to live a better life. Meanwhile, it aims to advance the digital transformation of the retail of goods and services on both the demand and supply sides.

Officially listed on the Main Board of the Stock Exchange of Hong Kong Limited on September 20, 2018, Meituan has always centered on customers and kept increasing investment in scientific R&D, thus better fulfilling its social responsibilities, creating more values for society, and seeking win-win cooperation with all partners.

About Intel

Intel (NASDAQ: INTC) is an industry leader, creating world-changing technology that enables global progress and enriches lives. Inspired by Moore's Law, we continuously work to advance the design and manufacturing of semiconductors to help address our customers' greatest challenges. By embedding intelligence in the cloud, network, edge and every kind of computing device, we unleash the potential of data to transform business and society for the better. To learn more about Intel's innovations, go to newsroom.intel.com and intel.com.

intel® 美团

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

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

Your costs and results may vary.

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

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

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