Case Study

Cloud Service Providers
Machine Learning

Low-latency Machine-learning Inference on Industry-standard Servers

Seeking improved cost efficiency, resource utilization and power consumption, NAVER and Intel collaborated to showcase the advantages of CPU-based machine-learning inference using software optimized for Intel® architecture.

At a Glance

- Intel® hardware and software can enable NAVER to run machine-learning inference workloads on industry-standard servers for greater data center efficiency
- Intel® Xeon® Scalable processors offer built-in AI capabilities that can enhance machine-learning performance
- NAVER can choose between Intel architecture-optimized machine-learning frameworks, depending on its needs
- The performance boost can range from up to 4.01x to 6.37x, depending on which optimized framework is used[2]

NAVER

South Korean cloud service provider (CSP) NAVER ranks 9th in Forbes' list of most innovative companies, and is number six in the list of Fortune 50 companies.[1] NAVER's messenger app boasts over 600 million users—25 million of which are in the U.S.—and its online search engine handles almost three-quarters of all web searches in South Korea. As NAVER expands its services, including artificial intelligence (AI) as a service, it is turning to Intel to help fine-tune its machine-learning hardware and software stack.

Challenge

NAVER puts its AI platform to use in several areas, one of which is identifying text in images. While performance is key to meeting service-level agreements (SLAs), cost efficiency, resource utilization and power consumption are also important considerations for the company's competitive edge. Using graphics processing unit (GPU)-based servers for inference provided great performance, but consumed too much power and left the company's CPU-based servers standing idle.

Solution

Working with NAVER, Intel tests showed the performance benefits that result from combining 2nd Generation Intel® Xeon® Scalable processors with a machine-learning framework optimized for Intel® architecture. The optimization tests showcased both PyTorch with Intel® PyTorch Extensions (IPEX), which was the easiest to deploy, and the Intel® Distribution of OpenVINO™ toolkit , which required more effort to deploy but delivered an additional performance boost. The test solutions would enable NAVER to meet its SLAs while allowing the company to get more work out of its CPU-based servers—driving up data center efficiency.

Results

Intel's tests included several configurations. The baseline configuration used the out-of-the-box version of PyTorch. This baseline was compared to three other configurations. The second configuration used PyTorch plus IPEX. The third configuration applied PyTorch Just-In-Time (JIT) functionality. The fourth configuration used the Intel Distribution of OpenVINO toolkit instead of PyTorch. Depending on the configuration, the normalized speed of inference increased between 4x and 6x.[2]

Achieving Low Latency and Data Center Efficiency

To keep its competitive edge, NAVER seeks to deliver low-latency AI services, such as with its Character-Region Awareness For Text detection (CRAFT) model. As part of NAVER's CLOVA platform, this service uses optical character recognition (OCR) to find the location of letters in an image and identify the type of letters from various languages and styles. While GPU-based servers provide low latency, they can also often drive up data center cooling costs as well. NAVER was looking for an alternative solution that could maintain its target latency while improving resource utilization and lowering total costs.

Software Optimization Is Key to Performance

In March 2020, the NAVER CLOVA OCR team and Intel began to discuss the potential for using Intel Xeon Scalable processors, combined with machine-learning frameworks optimized for Intel architecture, for the CRAFT model. NAVER was intrigued with the possibility of being able to meet latency SLAs with CPU-based servers. The two companies continued to communicate during the spring and summer, sharing information about NAVER's workload requirements and Intel's technology. The CLOVA team had previously worked with Intel engineers to use Intel technology to bolster CLOVA's natural language processing (NLP) for chatbot capabilities, so the teams had a close and productive relationship from the beginning.

After a few months, NAVER and Intel engaged in a proof of concept to explore how model inference performance is affected by optimizing the machine-learning model for Intel architecture. The baseline configuration ran the default version of PyTorch on a 2nd Gen Intel Xeon Scalable processor. Next, Intel demonstrated the benefits of using IPEX, which optimizes the Python package to take advantage of Intel hardware features such as Intel® Advanced Vector Extensions 512 (Intel® AVX-512). Performance was further increased by applying PyTorch JIT functionality, a method to create optimal graph models. With the help of JIT functionality, model parameters (such as weights) were cached to ensure low-latency data access during inference. Intel's tests also demonstrated how the Intel Distribution of OpenVINO toolkit could further improve inferencing speed. This toolkit provides developers with improved neural network performance on a variety of Intel® processors and helps them further unlock cost-effective, real-time vision applications. To demonstrate the scale-out ability of IPEX and OpenVINO on Intel architecture, the team tested both a one-socket and a two-socket system.

Based on Intel's tests, NAVER has the option of using one of several optimized software configurations to achieve the desired latency while improving data center efficiency.

Solution Ingredients

- Intel® Xeon® Gold 6240R processor
- Intel® Advanced Vector Extensions 512 (Intel® AVX-512)
- Intel® PyTorch extensions (IPEX) and PyTorch Just-in-Time (JIT) functionality
- Intel® Distribution of OpenVINO™ toolkit (optional)

Increase Performance up to 6x Depending on the Optimized Framework

Figure 1 illustrates the performance benefits that optimizing software for Intel hardware can bring. For example, adding IPEX to the default version of PyTorch resulted in up to 4.01x speedup, and using PyTorch JIT functionality increased that speedup to as much as 6.25x compared to just the default PyTorch. These performance gains are an "easy button"— they are available without any changes to existing code, and IPEX can be downloaded and installed from GitHub at no charge.

> "We look forward to continued collaboration, working closely with Intel to optimize our AI models and exploring other data types and Intel® Deep Learning Boost."
>
> **—Bado Lee,**
> OCR Leader, NAVER Corporation

Organizations looking for even more performance can use the Intel Distribution of OpenVINO toolkit. Although using this toolkit requires model conversion to take advantage of the optimizations, the resulting workload acceleration—up to 6.37x—can be a service differentiator for latency-sensitive use cases. The tests also showed that performance scales linearly as sockets are added.

Based on the proof of concept test results, Intel and NAVER will continue to explore how they can work together to transform NAVER's AI services and boost its competitive advantage in the AI as a service market.

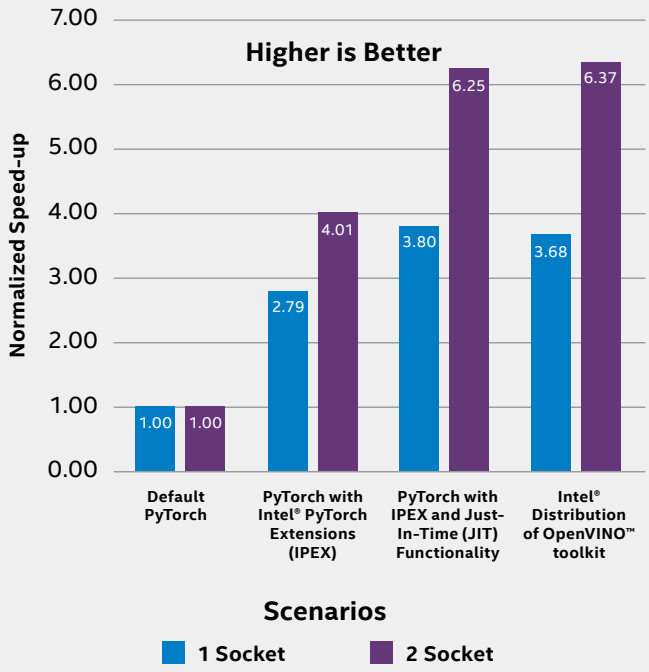

Figure 1. Compared to the default version of PyTorch, using Intel® architecture-optimized machine-learning frameworks can significantly boost inference performance.[2]

Spotlight on NAVER Corp. and the CLOVA Platform

Headquartered in Seongnam, South Korea, cloud service provider (CSP) NAVER introduced its artificial intelligence (AI) platform CLOVA (short for "cloud virtual assistant") in March 2017. CLOVA is an intelligent personal assistant for Android and iOS operating systems. The CLOVA platform first powered Wave, a smart speaker imbued with voice recognition capabilities. Over time, NAVER expanded the CLOVA platform to include other AI use cases such as image classification, optical character recognition (OCR), dubbing (adding voice to video) and smart home applications.

NAVER can also provide customized solutions that meet the needs of individual businesses by using data accumulated through NAVER's (and its subsidiary Line's) services and AI engines produced with CLOVA's own technology.

Learn More

You may find the following resources helpful:

- **Intel® Xeon® Scalable processors**
- **Intel® PyTorch Extensions**
- **Intel® Distribution of OpenVINO™ toolkit**
- **NAVER CLOVA AI Platform**

Find the solution that is right for your organization. Learn more about **Intel's software optimization for AI workloads**.

[1] https://www.prnewswire.com/in/news-releases/naver-leading-online-search-platform-in-south-korea-and-creator-of-the-line-messenger-app-deploys-bright-pattern-contact-center-software-821817064.html

[2] Tested by Intel as of 3/17/2021. 2-socket Intel® Xeon® Gold 6240R processor, 24 cores, Intel® Hyper-Threading Technology ON Intel® Turbo Boost Technology ON, Total Memory 192 GB (12 slots/16 GB/2666 MHz), BIOS: SE5C620.86B.02.01.0008.031920191559 (ucode: 0x5003003), Ubuntu 18.04 kernel 4.15.0-135-generic, Compiler: gcc 7.5.0, Deep Learning Framework: PyTorch v1.7.0 + torch_ipex-1.2.0 (commit 593f4b921b5f1a1fae26b922b394cc63b79fd40d), oneDNN v1.8.0 (commit 2c8d20640d5068e2d85e378b266644fe86220e84), Craft-ResNet50: Code modified from https://github.com/clovaai/CRAFT-pytorch, batch size = 1, synthetic data, 1 instance/2 socket, data type = FP32.

Config 1: Baseline Deep Learning Framework: PyTorch v1.7.0

Config 2: PyTorch v1.7.0 + torch_ipex-1.2.0

Config 3: PyTorch v1.7.0 + torch_ipex-1.2.0 + with PyTorch JIT tracing

Config 4: openvino_2021.2.185