May 25, 2022

Intel oneDNN AI Optimizations Enabled as Default in TensorFlow

Intel and Google team up to enable the oneDNN library as the default backend CPU optimization for TensorFlow 2.9.

SANTA CLARA, Calif.--(BUSINESS WIRE)-- **What's New:** In the latest release of TensorFlow 2.9, the performance improvements delivered by theIntel® oneAPI Deep Neural Network Library (oneDNN) are turned on by default. This applies to all Linux x86 packages and for CPUs with neural-network-focused hardware features (like AVX512_VNNI, AVX512_BF16, and AMX vector and matrix extensions that maximize AI performance through efficient compute resource usage, improved cache utilization and efficient numeric formatting) found on 2nd Gen Intel® Xeon® Scalable processors and newer CPUs. These optimizations enabled by oneDNN accelerate key performance-intensive operations such as convolution, matrix multiplication and batch normalization, with up to 3 times performance improvements compared to versions without oneDNN acceleration.

"Thanks to the years of close engineering collaboration between Intel and Google, optimizations in the oneDNN library are now default for x86 CPU packages in TensorFlow. This brings significant performance acceleration to the work of millions of TensorFlow developers without the need for them to change any of their code. This is a critical step to deliver faster AI inference and training and will help drive AI Everywhere."

–Wei Li, Intel vice president and general manager of AI and Analytics

Why It's Important: oneDNN performance improvements becoming available by default in the official TensorFlow 2.9 release will enable millions of developers who already use TensorFlow to seamlessly benefit from Intel software acceleration, leading to productivity gains, faster time to train and efficient utilization of compute. Additional TensorFlow-based applications, including TensorFlow Extended, TensorFlow Hub and TensorFlow Serving also have the oneDNN optimizations. TensorFlow has included experimental support for oneDNN since TensorFlow 2.5.

oneDNN is an open source cross-platform performance library of basic deep learning building blocks intended for developers of deep learning applications and frameworks. The applications and frameworks that are enabled by it can then be used by deep learning practitioners. oneDNN is part of oneAPI, an open, standards-based, unified programming model for use across CPUs as well as GPUs and other AI accelerators.

While there is an emphasis placed on AI accelerators like GPUs for machine learning and, in particular, deep learning, CPUs continue to play a large role across all stages of the AI workflow. Intel's extensive software-enabling work makes AI frameworks, such as the TensorFlow platform, and a wide range of AI applications run faster on Intel hardware that is ubiquitous across most personal devices, workstations and data centers. Intel's rich portfolio

of optimized libraries, frameworks and tools serves end-to-end AI development and deployment needs while being built on the foundation of oneAPI.

What This Helps Enable: The oneDNN-driven accelerations to TensorFlow deliver remarkable performance gains that benefit applications spanning natural language processing, image and object recognition, autonomous vehicles, fraud detection, medical diagnosis and treatment and others.

Deep learning and machine learning applications have exploded in number due to increases in processing power, data availability and advanced algorithms. TensorFlow has been one of the world's most popular platforms for AI application development with over 100 million downloads. Intel-optimized TensorFlow is available both as a standalone component and through the Intel® oneAPI AI Analytics Toolkit, and is already being used across a broad range of industry applications including the Google Health project, animation filmmaking at Laika Studios, language translation at Lilt, natural language processing at IBM Watson and many others.

More Context: Intel AI Software Tools | Intel AI | TensorFlow | oneAPI | oneDNN

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.

Notices and Disclaimers

Performance varies by use, configuration and other factors. Learn more at www.intel.com/PerformanceIndex. Results may vary.

Performance results are based on testing as of dates shown in configurations and may not reflect all publicly available updates.

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

View source version on businesswire.com:
https://www.businesswire.com/news/home/20220524005460/en/

Lindsey Barber-Montalvo
1-312-228-6875
lindsey.barber@ketchum.com

Source: Intel