

NVIDIA Expands Its Deep Learning Inference Capabilities for Hyperscale Datacenters

Company Unveils NVIDIA TensorRT 4, TensorFlow Integration, Kaldi Speech Acceleration and Expanded ONNX Support; GPU Inference Now up to 190x Faster Than CPUs

GPU Technology Conference -- NVIDIA today announced a series of new technologies and partnerships that expand its potential inference market to 30 million hyperscale servers worldwide, while dramatically lowering the cost of delivering deep learning-powered services.

Speaking at the opening keynote of GTC 2018, NVIDIA founder and CEO Jensen Huang described how GPU acceleration for deep learning inference is gaining traction, with new support for capabilities such as speech recognition, natural language processing, recommender systems, and image recognition -- in datacenters and automotive applications, as well as in embedded devices like robots and drones.

NVIDIA announced a new version of its TensorRT inference software, and the integration of TensorRT into Google's popular TensorFlow framework. NVIDIA also announced that Kaldi, the most popular framework for speech recognition, is now optimized for GPUs. NVIDIA's close collaboration with partners such as Amazon, Facebook and Microsoft make it easier for developers to take advantage of GPU acceleration using ONNX and WinML.

"GPU acceleration for production deep learning inference enables even the largest neural networks to be run in real time and at the lowest cost," said Ian Buck, vice president and general manager of Accelerated Computing at NVIDIA. "With rapidly expanding support for more intelligent applications and frameworks, we can now improve the quality of deep learning and help reduce the cost for 30 million hyperscale servers."

TensorRT, TensorFlow Integration

NVIDIA unveiled [TensorRT 4 software](#) to accelerate deep learning inference across a broad range of applications. TensorRT offers highly accurate INT8 and FP16 network execution, which can cut datacenter costs by up to 70 percent.⁽¹⁾

TensorRT 4 can be used to rapidly optimize, validate and deploy trained neural networks in hyperscale datacenters, embedded and automotive GPU platforms. The software delivers up to 190x⁽²⁾ faster deep learning inference compared with CPUs for common applications such as computer vision, neural machine translation, automatic speech recognition, speech synthesis and recommendation systems.

To further streamline development, NVIDIA and Google engineers have integrated TensorRT into TensorFlow 1.7, making it easier to run deep learning inference applications on GPUs.

Rajat Monga, engineering director at Google, said, "The TensorFlow team is collaborating very closely with NVIDIA to bring the best performance possible on NVIDIA GPUs to the deep learning community. TensorFlow's integration with NVIDIA TensorRT now delivers up to 8x higher inference throughput (compared to regular GPU execution within a low-latency target) on NVIDIA deep learning platforms with Volta Tensor Core technology, enabling the highest performance for GPU inference within TensorFlow."

NVIDIA has optimized the world's leading speech framework, Kaldi, to achieve faster performance running on GPUs. GPU speech acceleration will mean more accurate and useful virtual assistants for consumers, and lower deployment costs for datacenter operators.

Broad Industry Support

Developers at a wide spectrum of companies around the world are using TensorRT to discover new insights from data and to deploy intelligent services to businesses and consumers.

NVIDIA engineers have worked closely with Amazon, Facebook and Microsoft to ensure developers using ONNX frameworks such as Caffe 2, Chainer, CNTK, MXNet and Pytorch can now easily deploy to NVIDIA deep learning platforms.

Markus Noga, head of Machine Learning at SAP, said, "In our evaluation of TensorRT running our deep learning-based recommendation application on NVIDIA Tesla V100 GPUs, we experienced a 45x increase in inference speed and throughput compared with a CPU-based platform. We believe TensorRT could dramatically improve productivity for our enterprise customers."

Nicolas Koumchatzky, head of Twitter Cortex, said, "Using GPUs made it possible to enable media understanding on our platform, not just by drastically reducing media deep learning models training time, but also by allowing us to derive real-time understanding of live videos at inference time."

Microsoft also recently announced AI support for Windows 10 applications. NVIDIA partnered with Microsoft to build GPU-accelerated tools to help developers incorporate more intelligent features in Windows applications.

NVIDIA also announced GPU acceleration for Kubernetes to facilitate enterprise inference deployment on multi-cloud GPU clusters. NVIDIA is contributing GPU enhancements to the open-source community to support the Kubernetes ecosystem.

In addition, MathWorks, makers of MATLAB software, today announced TensorRT integration with MATLAB. Engineers and scientists can now automatically generate high-performance inference engines from MATLAB for the [NVIDIA DRIVE](#)™, [Jetson](#)™, and [Tesla](#)® platforms.

Inference for the Datacenter

Datacenter managers constantly balance performance and efficiency to keep their server fleets at maximum productivity. NVIDIA Tesla GPU-accelerated servers can replace several racks of CPU servers for deep learning inference applications and services, freeing up precious rack space and reducing energy and cooling requirements.

Inference for Self-Driving Cars, Embedded

TensorRT can also be deployed on NVIDIA DRIVE autonomous vehicles and NVIDIA Jetson embedded platforms. Deep neural networks on every framework

can be trained on [NVIDIA DGX™ systems](#) in the datacenter, and then deployed into all types of devices -- from robots to autonomous vehicles -- for real-time inferencing at the edge.

With TensorRT, developers can focus on developing novel deep learning-powered applications rather than performance tuning for inference deployment. Developers can use TensorRT to deliver lightning-fast inference using INT8 or FP16 precision that significantly reduces latency, which is vital for capabilities like object detection and path planning on embedded and automotive platforms.

Members of the [NVIDIA Developer Program](#) can learn more about the TensorRT 4 Release Candidate at <https://developer.nvidia.com/tensorrt>.

About NVIDIA

[NVIDIA](#)'s (NASDAQ:[NVDA](#)) invention of the GPU in 1999 sparked the growth of the PC gaming market, redefined modern computer graphics and revolutionized parallel computing. More recently, GPU deep learning ignited modern AI — the next era of computing — with the GPU acting as the brain of computers, robots and self-driving cars that can perceive and understand the world. More information at <http://nvidianews.nvidia.com/>.

1. Total cost of ownership based on a workload mix representative of a major cloud service provider: 60 percent neural collaborative filtering (NCF), 20 percent neural machine translation (NMT), 15 percent automatic speech recognition (ASR), 5 percent computer vision (CV) and per socket (Tesla V100 GPU vs CPU) workload speedups of: 10x NCF, 20x NMT, 15x ASR, 40x CV. CPU node configuration is two-socket Intel Skylake 6130. GPU recommended node configuration is the eight Volta HGX-1.
2. Performance gains observed over a range of important workloads. Examples include ResNet50 v1 inference performance at a 7 ms latency is 190x faster with TensorRT on a Tesla V100 GPU than using TensorFlow on a single-socket Intel Skylake 6140 at minimum latency (batch = 1).

Certain statements in this press release including, but not limited to, statements as to: the benefits, impact, performance, uses and abilities of NVIDIA TensorRT 4 and its integration into the TensorFlow framework; GPU acceleration for deep learning inference gaining traction and its impact and benefits; the benefits and impact of NVIDIA working with Kaldi to optimize GPUs, NVIDIA collaborating with partners, TensorRT being used around the world, GPU acceleration for Kubernetes, NVIDIA Tesla GPU-accelerated servers and MATLAB software's integration of TensorRT; NVIDIA's technologies expanding its potential inference market; and the support for intelligent applications and frameworks improving the quality of deep learning and reducing costs of hyperscale servers are forward-looking statements that are subject to risks and uncertainties that could cause results to be materially different than expectations. Important factors that could cause actual results to differ materially include: global economic conditions; our reliance on third parties to manufacture, assemble, package and test our products; the impact of technological development and competition; development of new products and technologies or enhancements to our existing product and technologies; market acceptance of our products or our partners' products; design, manufacturing or software defects; changes in consumer preferences or demands; changes in industry standards and interfaces; unexpected loss of performance of our products or technologies when integrated into systems; as well as other factors detailed from time to time in the reports NVIDIA files with the Securities and Exchange Commission, or SEC, including its Form 10-K for the fiscal period ended January 28, 2018. Copies of reports filed with the SEC are posted on the company's website and are available from NVIDIA without charge. These forward-looking statements are not guarantees of future performance and speak only as of the date hereof, and, except as required by law, NVIDIA disclaims any obligation to update these forward-looking statements to reflect future events or circumstances.

© 2018 NVIDIA Corporation. All rights reserved. NVIDIA, the NVIDIA logo, NVIDIA DGX, NVIDIA DRIVE, Jetson and Tesla are trademarks and/or registered trademarks of NVIDIA Corporation in the U.S. and other countries. Other company and product names may be trademarks of the respective companies with which they are associated. Features, pricing, availability and specifications are subject to change without notice.

Media Contacts

Kristin Bryson
+1 203 241 9190
kbryson@nvidia.com