NVIDIA Inference Breakthrough Makes Conversational AI Smarter, More Interactive From Cloud to Edge

TensorRT 8 Provides Leading Enterprises Across Healthcare, Automotive, Finance with World's Fastest AI Inference Performance

NVIDIA today launched TensorRT™ 8, the eighth generation of the company's AI software, which slashes inference time in half for language queries -- enabling developers to build the world's best-performing search engines, ad recommendations and chatbots and offer them from the cloud to the edge.

TensorRT 8's optimizations deliver record-setting speed for language applications, running BERT-Large, one of the world's most widely used transformer-based models, in 1.2 milliseconds. In the past, companies had to reduce their model size, which resulted in significantly less accurate results. Now, with TensorRT 8, companies can double or triple their model size to achieve dramatic improvements in accuracy.

"AI models are growing exponentially more complex, and worldwide demand is surging for real-time applications that use AI. That makes it imperative for enterprises to deploy state-of-the-art inferencing solutions," said Greg Estes, vice president of developer programs at NVIDIA. "The latest version of TensorRT introduces new capabilities that enable companies to deliver conversational AI applications to their customers with a level of quality and responsiveness that was never before possible."

In five years, more than 350,000 developers across 27,500 companies in wide-ranging areas, including healthcare, automotive, finance and retail, have downloaded TensorRT nearly 2.5 million times. TensorRT applications can be deployed in hyperscale data centers, embedded or automotive product platforms.

Latest Inference Innovations
In addition to transformer optimizations, TensorRT 8's breakthroughs in AI inference are made possible through two other key features.

Sparsity is a new performance technique in NVIDIA Ampere architecture GPUs to increase efficiency, allowing developers to accelerate their neural networks by reducing computational operations.

Quantization aware training enables developers to use trained models to run inference in INT8 precision without losing accuracy. This significantly reduces compute and storage overhead for efficient inference on Tensor Cores.

Broad Industry Support
Industry leaders have embraced TensorRT for their deep learning inference applications in conversational AI and across a range of other fields.

Hugging Face is an open-source AI leader relied on by the world's largest AI service providers across multiple industries. The company is working closely with NVIDIA to introduce groundbreaking AI services that enable text analysis, neural search and conversational applications at scale.

"We're closely collaborating with NVIDIA to deliver the best possible performance for state-of-the-art models on NVIDIA GPUs," said Jeff Boudier, product director at Hugging Face. "The Hugging Face Accelerated Inference API already delivers up to 100x speedup for transformer models powered by NVIDIA GPUs. With TensorRT 8, Hugging Face achieved 1ms inference latency on BERT, and we're excited to offer this performance to our customers later this year."

GE Healthcare, a leading global medical technology, diagnostics and digital solutions innovator, is using TensorRT to help

accelerate computer vision applications for ultrasounds, a critical tool for the early detection of diseases. This enables clinicians to deliver the highest quality of care through its intelligent healthcare solutions.

“When it comes to ultrasound, clinicians spend valuable time selecting and measuring images. During the R&D project leading up to the Vivid Patient Care Elevated Release, we wanted to make the process more efficient by implementing automated cardiac view detection on our Vivid E95 scanner,” said Erik Steen, chief engineer of Cardiovascular Ultrasound at GE Healthcare. “The cardiac view recognition algorithm selects appropriate images for analysis of cardiac wall motion. TensorRT, with its real-time inference capabilities, improves the performance of the view detection algorithm and it also shortened our time to market during the R&D project.”

Availability
TensorRT 8 is now generally available and free of charge to members of the NVIDIA Developer program. The latest versions of plug-ins, parsers and samples are also available as open source from the TensorRT GitHub repository.

About NVIDIA
NVIDIA’s (NASDAQ: NVDA) invention of the GPU in 1999 sparked the growth of the PC gaming market and has redefined modern computer graphics, high performance computing and artificial intelligence. The company’s pioneering work in accelerated computing and AI is reshaping trillion-dollar industries, such as transportation, healthcare and manufacturing, and fueling the growth of many others. More information at https://nvidianews.nvidia.com/.

Certain statements in this press release including, but not limited to, statements as to: the benefits, impact, performance, features, and availability of our products and services; worldwide demand for real-time applications that use AI surging; our collaborations with third parties; and industry leaders embracing TensorRT 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 most recent reports NVIDIA files with the Securities and Exchange Commission, or SEC, including, but not limited to, its annual report on Form 10-K and quarterly reports on Form 10-Q. 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.

Kristin Uchiyama
Enterprise and Edge Computing
+1-408-486-2248
kuchiyama@nvidia.com