NVIDIA Enables Era of Interactive Conversational AI with New Inference Software

NVIDIA TensorRT 7's Compiler Delivers Real-Time Inference for Smarter Human-to-AI Interactions

NVIDIA today introduced groundbreaking inference software that developers everywhere can use to deliver conversational AI applications, slashing inference latency that until now has impeded true, interactive engagement.

NVIDIA TensorRT™ 7 -- the seventh generation of the company's inference software development kit -- opens the door to smarter human-to-AI interactions, enabling real-time engagement with applications such as voice agents, chatbots and recommendation engines.

It is estimated that there are 3.25 billion digital voice assistants being used in devices around the world, according to Juniper Research. By 2023, that number is expected to reach 8 billion, more than the world's total population.

TensorRT 7 features a new deep learning compiler designed to automatically optimize and accelerate the increasingly complex recurrent and transformer-based neural networks needed for AI speech applications. This speeds the components of conversational AI by more than 10x compared to when run on CPUs, driving latency below the 300-millisecond threshold considered necessary for real-time interactions.

"We have entered a new chapter in AI, where machines are capable of understanding human language in real time," said NVIDIA founder and CEO Jensen Huang at his GTC China keynote. "TensorRT 7 helps make this possible, providing developers everywhere with the tools to build and deploy faster, smarter conversational AI services that allow more natural human-to-AI interaction."

Some of the world's largest, most innovative companies are already taking advantage of NVIDIA's conversational AI acceleration capabilities. Among these is Sogou, which provides search services to WeChat, the world's most frequently used application on mobile phones.

"Sogou provides high-quality AI services, such as voice, image, translation, dialogue and Q&A to hundreds of millions of users every day," said Yang Hongtao, CTO of Sogou. "By using the NVIDIA TensorRT inference platform, we enable online service responses in real time. These leading AI capabilities have significantly improved our user experience."

Rising Importance of Recurrent Neural Networks
TensorRT 7 speeds up a growing universe of AI models that are being used to make predictions on time-series, sequence-data scenarios that use recurrent loop structures, called RNNs. In addition to being used for conversational AI speech networks, RNNs help with arrival time planning for cars or satellites, prediction of events in electronic medical records, financial asset forecasting and fraud detection.

An explosion of combinations for RNN configurations and functions has created a challenge to rapidly deploy production code that meets real-time performance criteria -- causing months-long delays while developers created hand-written code optimizations. As a result, conversational AI has been limited to the few companies with the necessary talent.

With TensorRT's new deep learning compiler, developers everywhere now have the ability to automatically optimize these networks -- such as bespoke automatic speech recognition networks, and WaveRNN and Tacotron 2 for text-to-speech -- and to deliver the best possible performance and lowest latencies.

The new compiler also optimizes transformer-based models like BERT for natural language processing.

Accelerating Inference from Edge to Cloud
TensorRT 7 can rapidly optimize, validate and deploy a trained neural network for inference by hyperscale data centers, embedded or automotive GPU platforms.

NVIDIA's inference platform -- which includes TensorRT, as well as several NVIDIA CUDA-X AI™ libraries and NVIDIA GPUs -- delivers low-latency, high-throughput inference for applications beyond conversational AI, including image classification, fraud detection, segmentation, object detection and recommendation engines. Its capabilities are widely used by some of the world's leading enterprise and consumer technology companies, including Alibaba, American Express, Baidu, PayPal, Pinterest, Snap, Tencent and Twitter.

Availability
TensorRT 7 will be available in the coming days for development and deployment, without charge to members of the NVIDIA Developer program from the TensorRT webpage. 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, 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/.

Certain statements in this press release including, but not limited to, statements as to: the benefits, impact and availability of NVIDIA TensorRT 7 and NVIDIA's inference platform; the estimated number of digital voice assistants around the world and its expected growth; entering a new chapter in AI where machines are capable of understanding human language in real time; the impact of an explosion of combinations for RNN configurations and functions 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.

Media Contacts

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