NVIDIA Achieves Breakthroughs in Language Understanding to Enable Real-Time Conversational AI

Trains BERT in Record-Setting 53 Minutes and Slashes Inference to 2 Milliseconds; Enables Microsoft, Others to Use State-of-the-Art Language Understanding in Large-Scale Applications

NVIDIA today announced breakthroughs in language understanding that allow businesses to engage more naturally with customers using real-time conversational AI.

NVIDIA's AI platform is the first to train one of the most advanced AI language models -- BERT -- in less than an hour and complete AI inference in just over 2 milliseconds. This groundbreaking level of performance makes it possible for developers to use state-of-the-art language understanding for large-scale applications they can make available to hundreds of millions of consumers worldwide.

Early adopters of NVIDIA's performance advances include Microsoft and some of the world's most innovative startups, which are harnessing NVIDIA's platform to develop highly intuitive, immediately responsive language-based services for their customers.

Limited conversational AI services have existed for several years. But until this point, it has been extremely difficult for chatbots, intelligent personal assistants and search engines to operate with human-level comprehension due to the inability to deploy extremely large AI models in real time. NVIDIA has addressed this problem by adding key optimizations to its AI platform -- achieving speed records in AI training and inference and building the largest language model of its kind to date.

"Large language models are revolutionizing AI for natural language," said Bryan Catanzaro, vice president of Applied Deep Learning Research at NVIDIA. "They are helping us solve exceptionally difficult language problems, bringing us closer to the goal of truly conversational AI. NVIDIA's groundbreaking work accelerating these models allows organizations to create new, state-of-the-art services that can assist and delight their customers in ways never before imagined."

Fastest Training, Fastest Inference, Largest Model

AI services powered by natural language understanding are expected to grow exponentially in the coming years. Digital voice assistants alone are anticipated to climb from 2.5 billion to 8 billion within the next four years, according to Juniper Research. Additionally, Gartner predicts, by 2021, 15% of all customer service interactions will be completely handled by AI, an increase of 400% from 2017.[1]

Helping lead this new era, NVIDIA has fine-tuned its AI platform with key optimizations that have resulted in three new natural language understanding performance records:

- Fastest training: Running the large version of one of the world's most advanced AI language models -- Bidirectional Encoder Representations from Transformers (BERT) -- an NVIDIA DGX SuperPOD™ using 92 NVIDIA DGX-2H™ systems running 1,472 NVIDIA V100 GPUs slashed the typical training time for BERT-Large from several days to just 53 minutes. Additionally, NVIDIA trained BERT-Large on just one NVIDIA DGX-2 system in 2.8 days - demonstrating NVIDIA GPUs' scalability for conversational AI.
- Fastest inference: Using NVIDIA T4 GPUs running NVIDIA TensorRT™, NVIDIA performed inference on the BERT-Base SQuAD dataset in only 2.2 milliseconds - well under the 10-millisecond processing threshold for many real-time applications, and a sharp improvement from over 40 milliseconds measured with highly optimized CPU code.
- Largest model: With a focus on developers' ever-increasing need for larger models, NVIDIA Research built and trained the world's largest language model based on Transformers, the technology building block used for BERT and a growing number of other natural language AI models. NVIDIA's custom model, with 8.3 billion parameters, is 24 times the size of BERT-Large.

Ecosystem Adoption

Hundreds of developers worldwide are already using NVIDIA's AI platform to advance their own language understanding research and create new services.

Microsoft Bing is using the power of its Azure AI platform and NVIDIA technology to run BERT and drive more accurate search results.

"Microsoft Bing relies on the most advanced AI models and computing platform to deliver the best global search experience possible for our customers," said Rangan Majumder, group program manager, Microsoft Bing. "In close collaboration with NVIDIA, Bing further optimized the inferencing of the popular natural language model BERT using NVIDIA GPUs, part of Azure AI infrastructure, which led to the largest improvement in ranking search quality Bing deployed in the last year. We achieved two times the latency reduction and five times throughput improvement during inference using Azure NVIDIA GPUs compared with a CPU-based platform, enabling Bing to offer a more relevant, cost-effective, real-time search experience for all our customers globally."

Several startups in NVIDIA's Inception program, including Clinc, Passage AI and Recordsure, are also using NVIDIA's AI platform to build cutting-edge conversational AI services for banks, car manufacturers, retailers, healthcare providers, travel and hospitality companies, and more.

Clinc has made NVIDIA GPU-enabled conversational AI solutions accessible to more than 30 million people globally through a customer roster that includes leading car manufacturers, healthcare organizations and some of the world's leading financial institutions, including Barclays, USAA and Turkey's largest bank, Isbank.

"Clinc's leading AI platform understands complex questions and transforms them into powerful, actionable insights for the world's leading brands," said Jason Mars, CEO of Clinc. "The breakthrough performance that NVIDIA's AI platform provides has allowed us to push the boundaries of conversational AI and deliver revolutionary services that help our customers use technology to engage with their customers in powerful, more meaningful ways."

Optimizations Available Today

NVIDIA has made the software optimizations used to accomplish these breakthroughs in conversational AI available to developers:

- NVIDIA GitHub BERT training code with PyTorch *
- NGC model scripts and check-points for TensorFlow
- TensorRT optimized BERT Sample on GitHub
- Faster Transformer: C++ API, TensorRT plugin, and TensorFlow OP
- MXNet Gluon-NLP with AMP support for BERT (training and inference)
- TensorRT optimized BERT Jupyter notebook on AI Hub
- Megatron-LM: PyTorch code for training massive Transformer models

*NVIDIA's implementation of BERT is an optimized version of the popular Hugging Face repo

Additional Resources

- NVIDIA video: What's Next in Conversational AI
- NVIDIA Developer Blog: NVIDIA Clocks World's Fastest BERT Training Time and Largest Transformer Based Model Ever, Paving Path For Advanced Conversational AI
- NVIDIA Developer Blog: Real-Time Natural Language Understanding with BERT using TensorRT
- NVIDIA Applied Deep Learning Blog: MegatronLM: Training Billion+ Parameter Language Models Using GPU Model Parallelism

Keep Current on NVIDIA
Subscribe to the NVIDIA blog, follow us on Facebook, Twitter, LinkedIn and Instagram, and view NVIDIA videos on YouTube and images on Flickr.

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: NVIDIA achieving real-time breakthroughs in language understanding enabling real-time conversational AI and allowing businesses to engage more naturally with customers using real-time AI; the performance, impact and benefit of NVIDIA's technologies, NVIDIA's AI platform and BERT; NVIDIA's AI platform making it possible for developers to use language understanding for large-scale applications that they can make available to consumers worldwide; early adopters harnessing NVIDIA's platform to develop highly intuitive, responsive language-based services for their customers; NVIDIA achieving speed records in AI training and inference and building the largest language model of its kind to date; large language models revolutionizing AI for natural language, helping to solve language problems, and bringing us closer to conversational AI; NVIDIA's work accelerating models to create services that can assist customers in ways never before imagined; the expectation that AI services using natural language growing exponentially; the expected growth of digital voice assistants and the predicted growth of customer service interactions to be handled by AI; hundreds of developers using NVIDIA's AI platform to advance their research and create new services; Microsoft Bing using NVIDIA technology to run BERT and drive more accurate search results; NVIDIA and Microsoft collaborating to optimize Bing, and its benefits, impact and performance; startups using NVIDIA's AI platform to build cutting-edge AI services; NVIDIA's AI platform allowing Clinc to push the boundaries of conversation AI and deliver revolutionary services that help customers and engage them in new ways; and the availability of NVIDIA's code for BERT optimizations 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.

A video accompanying this announcement is available at:

https://www.globenewswire.com/NewsRoom/AttachmentNg/47a4da55-e11b-4cf2-9424-909941cf556b

Media Contacts

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