Nvidia launches TensorRT 8 for faster AI inference

: Dataquest ; Gurgaon (Jul 20, 2021).

ProQuest 文档链接

NVIDIA inference breakthrough makes conversational AI smarter, more interactive from cloud to edge.

NVIDIA 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.

"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."

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.

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.

Industry applications

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.”

:	Ultrasonic imaging
/:	: GE Healthcare; NAICS: 325412, 339112
/:	dqdeeptech
:	Nvidia launches TensorRT 8 for faster AI inference
:	Dataquest; Gurgaon
:	2021
:	Jul 20, 2021
:	Athena Information Solutions Pvt. Ltd.
:	Gurgaon
/:	India, Gurgaon
:	Computers--Data Base Management, Computers
ISSN:	0970034X
:	Trade Journal
:	English
:	News
ProQuest ID:	2553318397
URL:	https://www.proquest.com/trade-journals/nvidia-launches-tensorrt-8-faster-ai-inference/docview/2553318397/se-2?accountid=11524
:	Copyright 2021 Cyber Media (India) Ltd., distributed by Contify.com
:	2021-07-20
:	SciTech Premium Collection

CALIS e得文献获取, PQDT 全文库链接, UNICAT联合目录(刊名), 公共查询系统, Linking Service

条款与条件　联系 ProQuest