NVIDIA Jarvis Simplifies Building State-of-the-Art Conversational AI Services

New Application Framework Enables Creation of Custom, Language-Based AI Services from Customer Support to Real-Time Transcriptions

GTC 2020 -- NVIDIA today announced the release of NVIDIA Jarvis, a GPU-accelerated application framework that allows companies to use video and speech data to build state-of-the-art conversational AI services customized for their own industry, products and customers.

The shift toward working from home, telemedicine and remote learning has created a surge in demand for custom, language-based AI services, ranging from customer support to real-time transcriptions and summarization of video calls to keep people productive and connected.

Among the first companies to take advantage of Jarvis-based conversational AI products and services for their customers are Voca, an AI agent for call center support; Kensho, for automatic speech transcriptions for finance and business; and Square, with its virtual assistant for appointment scheduling.

“Conversational AI is central to the future of many industries, as applications gain the ability to understand and communicate with nuance and contextual awareness," said Jensen Huang, founder and CEO of NVIDIA. “NVIDIA Jarvis can help the healthcare, financial services, education and retail industries automate their overloaded customer support with speed and accuracy."

Applications built with Jarvis can take advantage of innovations in the new NVIDIA A100 Tensor Core GPU for AI computing and the latest optimizations in NVIDIA TensorRT™ for inference. For the first time, it's now possible to run an entire multimodal application, using the most powerful vision and speech models, faster than the 300-millisecond threshold for real-time interactions.

Jarvis provides a complete, GPU-accelerated software stack and tools making it easy for developers to create, deploy and run end-to-end, real-time conversational AI applications that can understand terminology unique to each company and its customers.

“IDC continues to see rapid growth within the conversational AI market largely because organizations of all sizes are beginning to realize the value of using well-trained virtual assistants and chatbots to help service their customers and grow their businesses," said David Schubmehl, research director of AI Software Platforms at IDC. “IDC expects worldwide spending on conversational AI use cases like automated customer service agents and digital assistants to grow from $5.8 billion in 2019 to $13.8 billion in 2023, a compound annual growth rate of 24 percent."

To offer an interactive, personalized experience, companies need to train their language-based applications on data that is specific to their own product offerings and customer requirements. However, building a service from scratch requires deep AI expertise, large amounts of data and compute resources to train the models, and software to regularly update models with new data.

Jarvis addresses these challenges by offering an end-to-end deep learning pipeline for conversational AI. It includes state-of-the-art deep learning models, such as NVIDIA's Megatron BERT for natural language understanding. Enterprises can further fine-tune these models on their data using NVIDIA NeMo, optimize for inference using TensorRT, and deploy in the cloud and at the edge using Helm charts available on NGC, NVIDIA's catalog of GPU-optimized software.

Early Adopters -- Voca, Kensho, Square

Companies worldwide are using NVIDIA's conversational AI platform to improve their services.

Voca's AI virtual agents -- which use NVIDIA for faster, more interactive, human-like engagements -- are used by Toshiba, AT&T and other world-leading companies. Voca uses AI to understand the full intent of a customer's spoken conversation and speech. This makes it possible for the agents to automatically identify different tones and vocal clues to discern between what a customer says and what a customer means. Additionally, using scalability features built into NVIDIA's AI platform, they can dramatically reduce customer wait time.

“Low latency is critical in call centers and with NVIDIA GPUs our agents are able to listen, understand and respond in under a second with the highest levels of accuracy," said Alan Bekker, co-founder and CTO of Voca. “Now our virtual agents are able to successfully handle 70-80 percent of all calls -- ranging from general customer service requests to payment transactions and technical support."

Kensho, the innovation hub for S&P Global located in Cambridge, Mass., that deploys scalable machine learning and analytics systems, has used NVIDIA's conversational AI to develop Scribe, a speech recognition solution for finance and business. With NVIDIA, Scribe outperforms other commercial solutions on earnings calls and similar financial audio in terms of accuracy by a margin of up to 20 percent.

“We're working closely with NVIDIA on ways to push end-to-end automatic speech recognition with deep learning even further," said Georg Kucsko, head of AI research at Kensho. “By training new models with NVIDIA, we're able to offer higher transcription accuracy for financial jargon compared to traditional approaches that do not use AI, offering our customers timely information in minutes versus days."

Square has created an AI virtual assistant that allows Square sellers to use AI to automatically confirm, cancel or change appointments with their customers, and free themselves to conduct more strategic customer engagement.

“Square Assistant can understand and provide help for 75 percent of customer questions, along with ensuring that 10 percent more people are showing up to their appointments," said Gabor Angeli, head of conversational AI at Square. “With GPUs, we're able to train models 10x faster versus CPUs to deliver more accurate, human-like interactions, ultimately helping our customers grow their businesses."

Availability
An early access program for NVIDIA Jarvis is available to a limited number of applicants. Developers interested in evaluating the application framework can sign up here.

Additional Resources

- NVIDIA Video: Using Conversational AI in Enterprise Applications
- Webinar: Training and Deploying Conversational AI Applications with NeMo and Jarvis
- NVIDIA Developer Blog: Introducing Jarvis: Framework for GPU-Accelerated Conversational AI Applications
- NVIDIA Developer Blog: Jumpstart Training for Speech Recognition Models in Different Languages with NeMo
- NVIDIA Developer Blog: NVIDIA NeMo: Fast Development of Speech and Language Models
- NVIDIA Developer Blog: State-of-the-Art Language Modeling Using Megatron on A100

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 https://nvidianews.nvidia.com/.

Certain statements in this press release including, but not limited to, statements as to: the benefits, performance and features of our products and technologies, including NVIDIA Jarvis, NVIDIA TensorRT, and NVIDIA A100 GPU; what Jarvis helps enable and allows companies to offer, including custom AI services and real-time transcriptions; the companies using Jarvis; conversational AI being central to the future of many industries; NVIDIA Jarvis enabling organizations to serve millions, improve customer satisfaction and support growing needs in industries; growth in the conversational AI market and its causes; the expectation for spending on conversational AI in the future; the requirements to offer language-based applications and how Jarvis addresses those challenges; enterprises being able to fine tune their models using NVIDIA products and technologies; how Voca, Kensho and Square use NVIDIA's platform and AI and its benefits and performance; and the availability of NVIDIA Jarvis 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.

© 2020 NVIDIA Corporation. All rights reserved. NVIDIA, the NVIDIA logo and TensorRT 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.

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