Publicado en Ciudad de México el 20/07/2021

# [El avance de NVIDIA Inference hace que la IA Conversacional sea más Inteligente e Interactiva](http://www.notasdeprensa.es)

## La IA Conversacional se hace más Inteligente e Interactiva desde la Nube hasta el Borde. TensorRT 8 ofrece a las empresas líderes en el sector de la Salud, la Automoción y las Finanzas el Rendimiento de Inferencia de IA más rápido del Mundo

 El día de hoy, NVIDIA lanzó TensorRT™ 8, la octava generación del revolucionario software de inferencia de la compañía, que reduce el tiempo de inferencia a la mitad para las consultas de idiomas, lo que permite a los desarrolladores crear los motores de búsqueda, las recomendaciones de anuncios y los chatbots con mejor rendimiento del mundo y ofrecerlos desde la nube hasta el borde and offer them from the cloud to the edge. Las optimizaciones de TensorRT 8 brindan una velocidad récord para aplicaciones de idiomas, ejecutando BERT-Large, uno de los modelos basados en transformadores más utilizados en el mundo, en 1.2 milisegundos. En el pasado, las empresas tenían que reducir el tamaño de su modelo, lo que generaba resultados significativamente menos precisos. Ahora, con TensorRT 8, las empresas pueden duplicar o triplicar el tamaño de su modelo para lograr mejoras dramáticas en la precisión. “Los modelos de IA se están volviendo exponencialmente más complejos, y la demanda mundial de aplicaciones en tiempo real que utilizan IA también está creciendo rápidamente, por lo que es imperativo que las empresas implementen soluciones de inferencia de vanguardia”, dijo Greg Estes, Vice-Presidente de Programas para Desarrolladores de NVIDIA. "La última versión de TensorRT presenta nuevas capacidades que permiten a las empresas ofrecer aplicaciones de IA conversacionales a sus clientes con un nivel de calidad y capacidad de respuesta que nunca antes había sido posible". En cinco años, más de 400,000 desarrolladores de 27,500 empresas en una amplia variedad de áreas, incluida la salud, la automoción, las finanzas y el comercio minorista, han descargado TensorRT casi 2.5 millones de veces. Las aplicaciones de TensorRT se pueden implementar en centros de datos de hiperescala, plataformas de productos integradas o automotrices. Últimas innovaciones de inferenciaAdemás de las optimizaciones del transformador, los avances de TensorRT 8 en la inferencia de IA son posibles gracias a otras dos características clave. Sparsity es una nueva técnica de rendimiento en las GPUs de Arquitectura NVIDIA Ampere para aumentar la eficiencia, lo que permite a los desarrolladores acelerar sus redes neuronales reduciendo las operaciones computacionales. La formación consciente de la cuantificación permite a los desarrolladores utilizar modelos entrenados para ejecutar inferencias con precisión INT8 sin perder precisión. Esto reduce significativamente la sobrecarga de procesamiento y almacenamiento para una inferencia eficiente en Tensor Cores. Amplio soporte en la industriaLos líderes de la industria han adoptado TensorRT para sus aplicaciones de inferencia de aprendizaje profundo en IA Conversacional y en una variedad de otros campos. Hugging Face es un código abierto líder en inteligencia artificial de en el que confían los proveedores de servicios de inteligencia artificial más grandes del mundo en múltiples industrias. La compañía está trabajando en estrecha colaboración con NVIDIA para presentar servicios de inteligencia artificial innovadores que permiten el análisis de texto, la búsqueda neuronal y las aplicaciones conversacionales a escala. "Estamos colaborando estrechamente con NVIDIA para ofrecer el mejor rendimiento posible para modelos de última generación en GPU NVIDIA", dijo Jeff Boudier, Director de Producto de Hugging Face. “La API de inferencia acelerada de cara abrazadora ya ofrece una aceleración de hasta 100 veces para los modelos de transformadores con tecnología de GPU NVIDIA. Con TensorRT 8, Hugging Face logró una latencia de inferencia de 1 ms en BERT, y estamos entusiasmados de ofrecer este rendimiento a nuestros clientes a finales de este año". GE Healthcare, un innovador líder mundial en tecnología médica, diagnóstico y soluciones digitales, está utilizando TensorRT para ayudar a acelerar las aplicaciones de visión por computadora para ultrasonidos, una herramienta crítica para la detección temprana de enfermedades. Esto permite a los médicos brindar la más alta calidad de atención a través de sus soluciones de atención médica inteligentes. “Cuando se trata de ultrasonidos, los médicos dedican un tiempo valioso a seleccionar y medir imágenes. Durante el proyecto de I + D que condujo a Vivid Patient Care Elevated Release, queríamos que el proceso fuera más eficiente implementando la detección automática de vista cardíaca en nuestro escáner Vivid E95 ”, dijo Erik Steen, Ingeniero en Jefe de Ultrasonido Cardiovascular en GE Healthcare. “El algoritmo de reconocimiento de la vista cardíaca selecciona las imágenes adecuadas para el análisis del movimiento de la pared cardíaca. TensorRT, con sus capacidades de inferencia en tiempo real, mejora el rendimiento del algoritmo de detección de vistas y también acorta nuestro tiempo de comercialización durante el proyecto de I + D". DisponibilidadTensorRT 8 ahora está disponible de forma general y gratuita para los miembros del Programa para Desarrolladores de NVIDIA. Las últimas versiones de complementos, analizadores y muestras también están disponibles como código abierto desde el repositorio de GitHub de TensorRT.

**Datos de contacto:**

Carlos Valencia

MKQ PR Agency

55 39 64 96 00

Nota de prensa publicada en: [https://www.mexicopress.com.mx/el-avance-de-nvidia-inference-hace-que-la-ia](http://www.notasdeprensa.es/educalivecom-disfruta-de-cursos-presenciales-y-clases-particulares-a-traves-de-internet)

Categorías: Inteligencia Artificial y Robótica Marketing Programación Hardware E-Commerce Ciudad de México



[**http://www.mexicopress.com.mx**](http://www.notasdeprensa.es)