AI Esperanto: Large Language Models Read Data With NVIDIA Triton

Companies bringing natural language processing to many markets are turning to Triton for AI inference.

Author: Rick Merritt

Julien Salinas wears many hats. He’s an entrepreneur, software developer and, until lately, a volunteer fireman in his mountain village an hour’s drive from Grenoble, a tech hub in southeast France.

He’s nurturing a two-year old startup, NLP Cloud , that’s already profitable, employs about a dozen people and serves customers around the globe. It’s one of many companies worldwide using NVIDIA software to deploy some of today’s most complex and powerful AI models.

NLP Cloud is an AI-powered software service for text data. A major European airline uses it to summarize internet news for its employees. A small healthcare company employs it to parse patient requests for prescription refills. An online app uses it to let kids talk to their favorite cartoon characters.

It’s all part of the magic of natural language processing (NLP), a popular form of AI that’s spawning some of the planet’s biggest neural networks called large language models . Trained with huge datasets on powerful systems, LLMs can handle all sorts of jobs such as recognizing and generating text with amazing accuracy.

NLP Cloud uses about 25 LLMs today, the largest has 20 billion parameters, a key measure of the sophistication of a model. And now it’s implementing BLOOM, an LLM with a whopping 176 billion parameters.

Running these massive models in production efficiently across multiple cloud services is hard work. That’s why Salinas turns to NVIDIA Triton Inference Server .

“Very quickly the main challenge we faced was server costs,” Salinas said, proud his self-funded startup has not taken any outside backing to date.

“Triton turned out to be a great way to make full use of the GPUs at our disposal,” he said.

For example, NVIDIA A100 Tensor Core GPUs can process as many as 10 requests at a time — twice the throughput of alternative software — thanks to FasterTransformer , a part of Triton that automates complex jobs like splitting up models across many GPUs.

FasterTransformer also helps NLP Cloud spread jobs that require more memory across multiple NVIDIA T4 GPUs while shaving the response time for the task.

Customers who demand the fastest response times can process 50 tokens — text elements like words or punctuation marks — in as little as half a second with Triton on an A100 GPU, about a third of the response time without Triton.

“That’s very cool,” said Salinas, who’s reviewed dozens of software tools on his personal blog.

Around the globe, other startups and established giants are using Triton to get the most out of LLMs.

Microsoft’s Translate service helped disaster workers understand Haitian Creole while responding to a 7.0 earthquake. It was one of many use cases for the service that got a 27x speedup using Triton to run inference on models with up to 5 billion parameters.

NLP provider Cohere was founded by one of the AI researchers who wrote the seminal paper that defined transformer models . It’s getting up to 4x speedups on inference using Triton on its custom LLMs, so users of customer support chatbots, for example, get swift responses to their queries.

NLP Cloud and Cohere are among many members of the NVIDIA Inception program, which nurtures cutting-edge startups. Several other Inception startups also use Triton for AI inference on LLMs.

Tokyo-based rinna created chatbots used by millions in Japan, as well as tools to let developers build custom chatbots and AI-powered characters. Triton helped the company achieve inference latency of less than two seconds on GPUs.

In Tel Aviv, Tabnine runs a service that's automated up to 30% of the code written by a million developers globally (see a demo below). Its service runs multiple LLMs on A100 GPUs with Triton to handle more than 20 programming languages and 15 code editors.

Twitter uses the LLM service of Writer , based in San Francisco. It ensures the social network's employees write in a voice that adheres to the company's style guide. Writer's service achieves a 3x lower latency and up to 4x greater throughput using Triton compared to prior software.

If you want to put a face to those words, Inception member Ex-human , just down the street from Writer, helps users create realistic avatars for games, chatbots and virtual reality applications. With Triton, it delivers response times of less than a second on an LLM with 6 billion parameters while reducing GPU memory consumption by a third.

It's another example of how LLMs are expanding AI's horizons .

Triton is widely used, in part, because its versatile. The software works with any style of inference and any AI framework — and it runs on CPUs as well as NVIDIA GPUs and other accelerators.

Back in France, NLP Cloud is now using other elements of the NVIDIA AI platform.

For inference on models running on a single GPU, it's adopting NVIDIA TensorRT software to minimize latency. "We're getting blazing-fast performance with it, and latency is really going down," Salinas said.

The company also started training custom versions of LLMs to support more languages and enhance efficiency. For that work, it's adopting NVIDIA Nemo Megatron , an end-to-end framework for training and deploying LLMs with trillions of parameters.

The 35-year-old Salinas has the energy of a 20-something for coding and growing his business. He describes plans to build private infrastructure to complement the four public cloud services the startup uses, as well as to expand into LLMs that handle speech and text-to-image to address applications like semantic search.

"I always loved coding, but being a good developer is not enough: You have to understand your customers' needs," said Salinas, who posted code on GitHub nearly 200 times last year.

If you're passionate about software, learn the latest on Triton in this technical blog .

Original URL: https://blogs.nvidia.com/blog/2022/10/05/ai-large-language-models-triton/