NeMo Megatron Reinforces NVIDIA AI Leadership in Large Language Models

Alberto Romero, LLM Analyst
Karl Freund, Founder and Principal Analyst
Cambrian-AI Research

Transformer-based large language models (LLMs) are reshaping the AI landscape today. Since OpenAI established the now generally accepted scaling laws of transformers with GPT-3 in 2020, AI companies have been exerting extreme effort to stay at the vanguard by scaling ever-larger models. NVIDIA has now demonstrated the company's NeMo Megatron as one of the most performant and efficient LLM platforms now available. Let's take a look at what NVIDIA has developed and what the competitive landscape presents.

NVIDIA: A LEADER FACING NEW OBSTACLES

It's been 5 years since Google invented the transformer architecture, which forms the basis for today's LLM renaissance. With the advent of this super versatile architecture—that is revamping the two main commercial AI branches; vision, and language—many new AI hardware companies emerged. They recognized the opportunities of AI-first approaches; building hardware specifically designed to handle transformer workloads. They could then leverage this singular edge against giants like NVIDIA.

While the AI industry awaits the newest NVIDIA Hopper GPU, which includes a dedicated engine for processing Transformers, Cerebras Systems has come forward claiming the largest LLM ever trained on a "single chip". This unicorn startup has been building extremely large chips as a full wafer-scale engine, or WSE. Its latest, the WSE-2 is 56x larger than an NVIDIA A-100 GPU and has more compute performance than a hundred of them—to manage enormous models several times larger than GPT-3 (175B parameters).

While this is certainly impressive, Cerebras shouldn't claim victory too soon. NVIDIA is counterattacking, first with NeMo Megatron and soon with Hopper. Of course as an established company with powerful clients around the world using its products and services, the company is a juggernaught in AI. NVIDIA only has to play its cards well to combat potential competitors which still need to build most of the infrastructure and a reliable full-stack surrounding their novel hardware ideas.

NVIDIA IS ADAPTING ITS AI PLATFORM TO LARGE LANGUAGE MODELS

One of the main limitations the company faces is that GPUs aren't specifically designed for transformer architectures either in training or inference, which reduces efficiency. One possible approach is to develop new hardware, as they've recently done with the H-100 transformer engine.

Another possibility that will improve efficiency for current generation GPUs is to optimize the software. That's what NVIDIA has tackled now. NVIDIA has announced updates to the NeMo Megatron framework that provides training speedups of up to 30%. This translates into a 175B-parameter model (e.g., GPT-3) being trained in 24 days instead of 34 using 1024 A-100 GPUs.

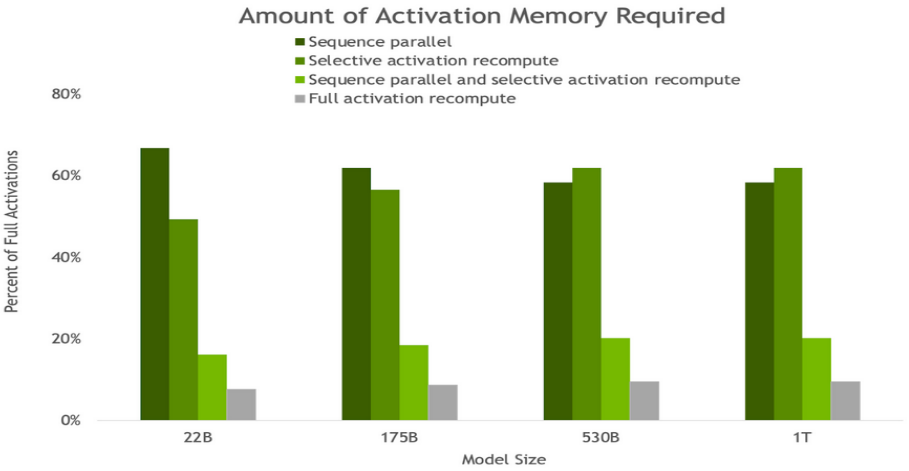

Figure 1: "Amount of activation memory saved in backward pass thanks to SP and SAR. As model size increases, both SP and SAR have similar memory savings, reducing the memory required by ~5x" - NVIDIA

One of the updates improves training efficiency. It comprises a couple of techniques to reduce memory requirements during training that reduces computational time. Sequence parallelism reduces activation memory requirements beyond the usual tensor and pipeline parallelism methods. The other technique, selective activation recomputation, is a novel approach that focuses on selecting activations with high-memory low-compute requirements to recompute when memory constraints are too tight— which avoids the inefficiency of full activation recomputation.

NVIDIA says the combination of both techniques provides a 5x reduction in memory requirements with respect to doing only tensor parallelism. The overhead is just 2–4% in comparison to +36% in the case of applying full activation recomputation. In total, this translates into a 30% improvement in throughput, and thus a significant reduction in training times across model sizes, from 22 billion to 1 trillion parameters.

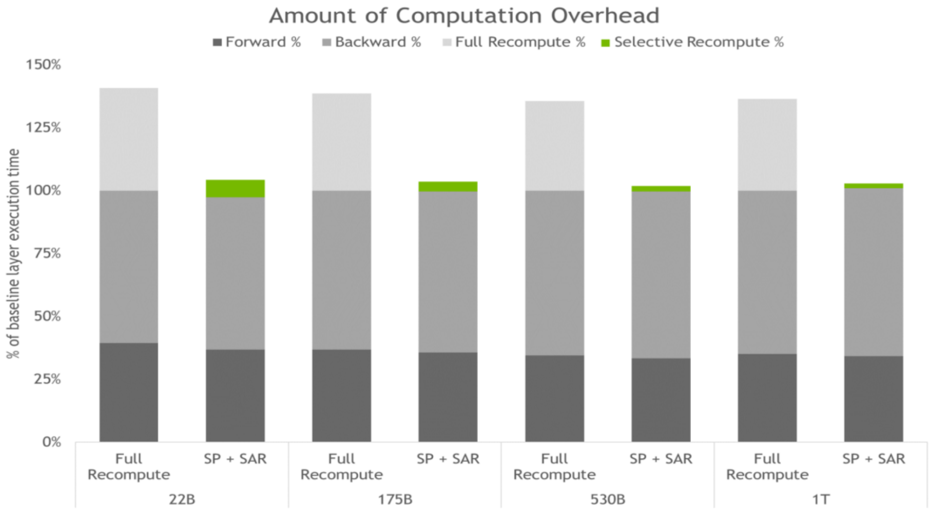

Figure 2: "Amount of computation overhead for full activation recomputation, and SP plus SAR. Bars represent per-layer breakdown of forward, backward, and recompute times. Baseline is the case with no recomputation and no sequence parallelism. These techniques are effective at reducing the overhead incurred when all activations are recomputed instead of saved. For the largest models, overhead drops from 36% to just 2%." -NVIDIA

The other update improves deployment times. It's a hyperparameter tool to "automatically find optimal training and inference configurations." Searching for the best configuration can be so costly that sometimes companies have to settle on an under-optimal solution using dubious heuristic techniques. However, using this tool, NVIDIA researchers found the "optimal training configuration for a 175B GPT-3 model in under 24 hours." This tool can also find the best trade-off between latency and throughput during inference to adapt to the user's needs.

Conclusions

The combination of these techniques makes NeMo Megatron better suited to handle transformer workloads across the entire stack. NVIDIA keeps finding solutions to improve its existing stack of products, services, and frameworks while continually improving the underlying hardware. We view these enhancements to NeMo as a prelude to what we can expect when the Hopper H100 GPU becomes available later this year.

NVIDIA is willing to work to keep its dominance regardless of how many changes occur in the AI field or how fast they happen. New companies emerge with singular ideas that may better fit current trends, but they still need to build the full stack that NVIDIA has been maturing for many years. That advantage is difficult to overcome; most startups dedicate at least 50% of their engineering staff to software. NVIDIA can simply adjust their stack to leverage the potential of strong trends while reinforcing its dominance where other companies can't—and probably won't—reach.

IMPORTANT INFORMATION ABOUT THIS PAPER

Authors: **Alberto Romero and Karl Freund, Cambrian-AI Research, LLC**

INQUIRIES:

Contact us if you would like to discuss this report, and Cambrian-AI Research will respond promptly.

CITATIONS

This paper can be cited by accredited press and analysts but must be cited in context, displaying the author's name, author's title, and "Cambrian-AI Research." Non-press and non-analysts must receive prior written permission from Cambrian-AI Research for any citations.

LICENSING

DISCLOSURES

This document was developed with QTI funding and support. Although the document may utilize publicly available material from various vendors, including QTI, it does not necessarily reflect the positions of such vendors on the issues addressed in this document.

DISCLAIMER

The information presented in this document is for informational purposes only and may contain technical inaccuracies, omissions, and typographical errors. Cambrian-AI Research disclaims all warranties as to the accuracy, completeness, or adequacy of such information and shall have no liability for errors, omissions, or inadequacies in such information. This document consists of the opinions of Cambrian-AI Research and should not be construed as statements of fact. The opinions expressed herein are subject to change without notice.

Cambrian-AI Research provides forecasts and forward-looking statements as directional indicators and not as precise predictions of future events. While our forecasts and forward-looking statements represent our current judgment on what the future holds, they are subject to risks and uncertainties that could cause actual results to differ materially. You are cautioned not to place undue reliance on these forecasts and forward-looking statements, which reflect our opinions only as of the date of publication for this document. Please keep in mind that we are not obligating ourselves to revise or publicly release the results of any revision to these forecasts and forward-looking statements in light of new information or future events.