Multiverse Computing Launches CompactifAI to Streamline LLMs to Reduce Energy Use and Compute Costs

Leveraging quantum-inspired tensor networks, the CompactifAI compressor can also accelerate AI data training and enable new portable use cases for LLMs

SAN SEBASTIÁN, SPAIN, November 15, 2023 /EINPresswire.com/ -- Multiverse Computing, a global leader in value-based quantum computing solutions, today announced CompactifAI, a solution for the high compute cost of machine learning algorithms. This new tool for managing the significant energy demands required to train and run large language models (LLMs) like ChatGPT and Bard is designed to reduce development costs and make it easier to integrate these models into more digital services.

> An innovative solution capable of optimizing use of resources holds the potential to enable the true scaling of LLM solutions."
>
> *Rafael San Juan, Global Innovation Manager at Iberdrola*

CompactifAI uses tensor networks to reduce the number of parameters in a model, reducing its overall size as well as shrinking memory and storage space requirements. Retraining has the potential to be faster as well since users can add new data to an existing model that has already been compressed to generate an updated model that retains the benefits of compression while preserving the quality of results.

CompactifAI is designed to reduce energy requirements at multiple points in the lifecycle of an LLM, including the training phase, general operations and retraining. Multiverse's new software also reduces the overall footprint of the models, making them more portable and easier to run at the edge in applications such as autonomous vehicles and remote production facilities. A video showing how CompactifAI works is available here.

"This new tool will lift a major barrier to the growth of large language models across industries: the sheer size of these algorithms and data sets and the energy required to run them," said Enrique Lizaso Olmos, CEO of Multiverse Computing. "CompactifAI also opens the door to new

LLM uses cases for on premise, on the edge and other applications where there is no dedicated cloud connection."

A new study of AI energy consumption found that by 2027 worldwide AI-related electricity consumption each year could match the annual energy use of countries such as the Netherlands, Argentina and Sweden.[1,2] The research conclusions are based on estimations of the amount of energy needed to run the most popular servers that power LLMs. The research also considered increased usage as the models become integrated into popular search engines and other elements of internet infrastructure.

LLMs already consume a significant amount of energy in both the training phase and in daily operations. A University of Washington researcher estimates that a single LLM could require up to 10 gigawatt-hours of power during the training phase, equal to the energy use of more than 1,000 U.S. households annually. To respond to hundreds of millions of queries each day, those same models could use up to 1 gigawatt-hour of power each day, which is roughly equal to the daily energy usage rate of 33,000 US households.[3]

"LLMs demand significant energy, compute power, and memory resources during their training process and during their lifecycle," said Rafael San Juan, Global Innovation Manager at Iberdrola, one of the world's largest electric utility providers and Multiverse customer. "An innovative solution capable of optimizing use of resources holds the potential to revolutionize the landscape, minimizing the impact this has on the electric grid and enabling the true scaling of LLM solutions."

The software offers low, medium and high options for compressing a model, depending upon the application requirements for a particular LLM. Multiverse expects AI developers to be the initial users for this software-as-a-service platform.

Originally used in the study of condensed matter physics, a tensor network is a visual language that describes complex systems. This visual language makes it easier to understand how each component in the system interacts with all the others and to predict the results of those interactions. Extracting information from tensor networks is relatively straightforward, providing another benefit to this approach to computation.

Multiverse Computing's Chief Science Officer Roman Orus was one of the first researchers to study tensor networks during a research fellowship at the University of Queensland in 2006. As a co-founder of the company, he applied this knowledge in one of Multiverse's first projects: using tensor networks for portfolio optimization with a multinational bank.[4]

More recently, researchers have used tensor networks as models for machine learning architecture and to compress the layers in neural networks.[5] Tensor networks can be mapped directly to quantum circuits as well. Researchers expect tensor networks to serve as a bridge between today's noisy quantum computers and the fault tolerant machines of the future.

Footnotes

[1] “The growing energy footprint of artificial intelligence,” Joule

[2] "Powering AI could use as much electricity as a small country,” Digiconomist - https://digiconomist.net/powering-ai-could-use-as-much-electricity-as-a-small-country/

[3] "Q&A: UW researcher discusses just how much energy ChatGPT uses," UW News, University of Washington - https://www.washington.edu/news/2023/07/27/how-much-energy-does-chatgpt-use/

[4] “BBVA pursues the financial sector’s ‘quantum advantage,’” BBVA - https://www.bbva.com/en/innovation/bbva-pursues-the-financial-sectors-quantum-advantage/

[5] “Applications of Tensor Networks to Machine Learning,” Tensor Network.org - https://tensornetwork.org/ml/

About Multiverse Computing

Multiverse Computing is a leading quantum software company that applies quantum and quantum-inspired solutions to tackle complex problems in finance, banking, manufacturing, energy, and cybersecurity to deliver value today and enable a more resilient and prosperous economy. The company’s expertise in quantum algorithms and quantum-inspired algorithms means it can secure maximum results from current quantum devices as well as classical high-performance computers. Its flagship product, Singularity, allows professionals across all industries to leverage quantum computing to speed up and improve the accuracy of optimization and AI models with existing and familiar software tools. The company also has developed CompactifAI, a compressor which uses quantum-inspired tensor networks to make AI systems such as large language models more efficient and portable. In addition to finance and AI, Multiverse serves enterprises in the mobility, energy, life sciences and industry 4.0 sectors. The company is based in San Sebastian, Spain, with branches in Toronto, Paris and Munich.

Contact Multiverse Computing at contact@multiversecomputing.com

Veronica Combs
HKA Marketing Communications
+1 714-422-0927
email us here

we have missed, please do bring it to our attention. Your help is welcome. EIN Presswire, Everyone's Internet News Presswire™, tries to define some of the boundaries that are reasonable in today's world. Please see our Editorial Guidelines for more information.