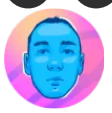
News Report Technology July 11, 2023

GPT-4's Leaked Details Shed Light on its Massive Scale and Impressive Architecture



by Damir Yalalov

Published: July 11, 2023 at 7:19 am Updated: July 11, 2023 at 7:23 am

by Danil Myakin

Edited and fact-checked: July 11, 2023 at 7:19 am

IN BRIEF

> The leaked information about GPT-4 has sparked excitement among the AI community. With over 10 times the parameters of its predecessor, GPT-3, GPT-4 is estimated to have 1.8 trillion parameters distributed across 120 layers.

OpenAI implemented a mixture of experts (MoE) model, utilizing 16 experts with 111 billion parameters for multi-layer perceptrons (MLP). The model's efficient inference process utilizes 280 billion parameters and

Hot Stories

The Red Ocean of the Derivatives Market: How Gate.io Stands Out
by Victoria d'Este January 01, 2025

Squid Game-Inspired Tokens Flood the Crypto Market as Investors Face Growing Scam Risks
by Victoria d'Este January 01, 2025

Vitalik Buterin Donates $170K in ETH to Tornado Cash Developers' Legal Fund
by Victoria d'Este January 01, 2025

Jordan Approves Blockchain Policy to Modernize Government Services
by Victoria d'Este January 01, 2025

Latest News

Vitalik Buterin Donates $170K in ETH to Tornado Cash Developers' Legal Fund
by Victoria d'Este January 01, 2025

Jordan Approves Blockchain Policy to Modernize Government Services
by Victoria d'Este January 01, 2025

Ether's 2025 Outlook: Challenges in Delivering Meaningful Rallies
by Victoria d'Este January 01, 2025

The US BTC Spot ETF Records Net Inflow of $5.3 Million
by Victoria d'Este January 01, 2025

from 8K to 32K.

> OpenAI utilized parallelism in GPT-4 to leverage the full potential of their A100 GPUs, employing 8-way tensor parallelism and 15-way pipeline parallelism. The training process was extensive and resource-intensive, with costs ranging from $32 million to $63 million.

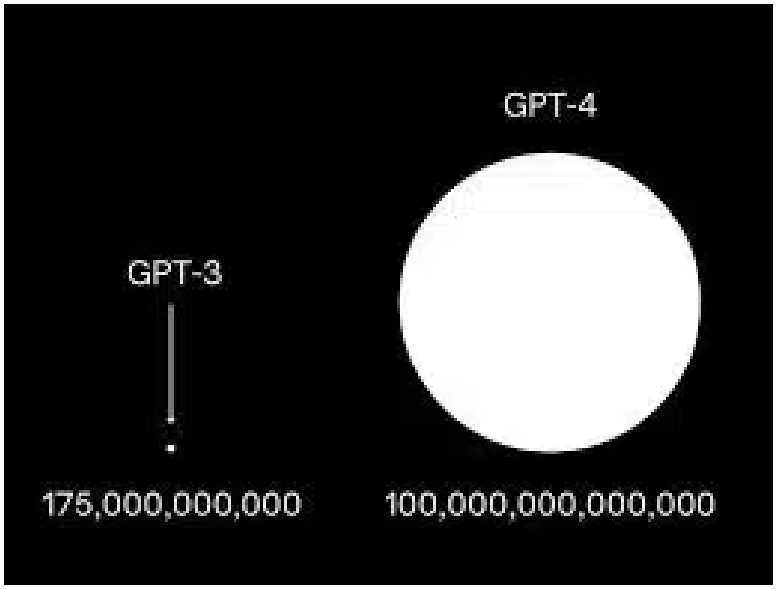
> GPT-4's inference cost is approximately three times higher than its predecessor, but it also incorporates multi-query attention, continuous batching, and speculative decoding. The inference architecture operates on a cluster of 128 GPUs, distributed across multiple data centers.

The recent leak of details surrounding GPT-4 has sent shockwaves through the AI community. The leaked information, obtained from an undisclosed source, provides a glimpse into the awe-inspiring capabilities and unprecedented scale of this groundbreaking model. We will break down the facts and unveil the key aspects that make GPT-4 a true technological marvel.

Credit: Metaverse Post (mpost.io)

GPT-4’s Massive Parameters Count

approximately 1.8 trillion parameters distributed across an impressive 120 layers. This substantial increase in scale undoubtedly contributes to GPT-4's enhanced capabilities and potential for groundbreaking advancements.



Mixture of Experts Model (MoE)

To ensure reasonable costs while maintaining exceptional performance, OpenAI implemented a mixture of experts (MoE) model in GPT-4. By utilizing 16 experts within the model, each consisting of around 111 billion parameters for multi-layer perceptrons (MLP), OpenAI effectively optimized resource allocation. Notably, during each forward pass, only two experts are routed, minimizing computational requirements without compromising results. This innovative approach demonstrates OpenAI's commitment to maximizing efficiency and cost-effectiveness in their models.

implications - by @dylan522p :

semianalysis.com/p/gpt-4-archit...

A non-paywalled summary can be found here: x.com/Yampeleg/statu...

6:38 PM · Jul 10, 2023

3 Reply Copy link

Read more on X

Simplified MoE Routing Algorithm

While the model often explores advanced routing algorithms for selecting experts to handle each token, OpenAI's approach in the current GPT-4 model is reportedly more straightforward. The routing algorithm employed by the AI is alleged to be relatively simple, but nonetheless effective. Approximately 55 billion shared parameters for attention facilitate the efficient distribution of tokens to the appropriate experts within the model.

Efficient Inference

GPT-4's inference process showcases its efficiency and computational prowess. Each forward pass, dedicated to generating a single token, utilizes approximately 280 billion parameters and 560 TFLOPs (tera floating-point operations per second). This stands in stark contrast to the immense scale of GPT-4, with its 1.8 trillion parameters and 3,700 TFLOPs per forward pass in a purely dense model. The efficient use of resources highlights OpenAI's

Dataset

GPT-4 has been trained on a colossal dataset comprising approximately 13 trillion tokens. It is important to note that these tokens include both unique tokens and tokens accounting for epoch numbers. The training process includes two epochs for text-based data and four epochs for code-based data. OpenAI leveraged millions of rows of instruction fine-tuning data sourced from ScaleAI and internally to refine the model's performance.

Refinement through Fine-Tuning from 8K to 32K

The pre-training phase of GPT-4 employed an 8k context length. Subsequently, the model underwent fine-tuning, resulting in the 32k version. This progression builds upon the pre-training phase, enhancing the model's capabilities and tailoring it to specific tasks.

Scaling with GPUs via Parallelism

OpenAI harnessed the power of parallelism in GPT-4 to leverage the full potential of their A100 GPUs. They employed 8-way tensor parallelism, which maximizes parallel processing, as it is the limit for NVLink. Additionally, 15-way pipeline parallelism was utilized to further enhance performance. While specific techniques such as ZeRo Stage 1 were likely employed, the exact methodology remains undisclosed.

intensive endeavor. OpenAI allocated approximately 25,000 A100 GPUs over a period of 90 to 100 days, operating at a utilization rate of approximately 32% to 36% MFU (most frequently used). The training process incurred numerous failures, necessitating frequent restarts from checkpoints. If estimated at $1 per A100 hour, the training costs for this run alone would amount to approximately $63 million.

Tradeoffs in Mixture of Experts

Implementing a mixture of experts model presents several tradeoffs. In the case of GPT-4, OpenAI opted for 16 experts instead of a higher number. This decision reflects a balance between achieving superior loss results and ensuring generalizability across various tasks. More experts can present challenges in terms of task generalization and convergence. OpenAI's choice to exercise caution in expert selection aligns with their commitment to reliable and robust performance.

Inference Cost

Compared to its predecessor, the 175 billion parameter Davinci model, GPT-4's inference cost is approximately three times higher. This discrepancy can be attributed to several factors, including the larger clusters required to support GPT-4 and the lower utilization achieved during inference. Estimations indicate an approximate cost of $0.0049 cents per 1,000 tokens for 128 A100 GPUs, and $0.0021 cents per 1,000 tokens for 128 H100 GPUs when inferring GPT-4 with an 8k. These figures assume decent

OpenAI leverages multi-query attention (MQA), a technique widely employed in the field, in GPT-4 as well. By implementing MQA, the model requires only one head, significantly reducing the memory capacity necessary for the key-value cache (KV cache). Despite this optimization, it should be noted that the 32k batch GPT-4 cannot be accommodated on 40GB A100 GPUs, and the 8k is constrained by the maximum batch size.

Continuous Batching

To strike a balance between latency and inference costs, OpenAI incorporates both variable batch sizes and continuous batching in GPT-4. This adaptive approach allows for flexible and efficient processing, optimizing resource utilization and reducing computational overhead.

Vision Multi-Modal

GPT-4 introduces a separate vision encoder alongside the text encoder, featuring cross-attention between the two. This architecture, reminiscent of Flamingo, adds additional parameters to the already impressive 1.8 trillion parameter count of GPT-4. The vision model undergoes separate fine-tuning using approximately 2 trillion tokens following the text-only pre-training phase. This vision capability empowers autonomous agents to read web pages, transcribe images, and interpret video content—an invaluable asset in the age of multimedia data.

Speculative Decoding

tokens are then fed into a larger "oracle" model as a single batch. If the smaller model's predictions align with the larger model's agreement, several tokens can be decoded together. However, if the larger model rejects the tokens predicted by the draft model, the rest of the batch is discarded, and inference continues solely with the larger model. This approach allows for efficient decoding while potentially accepting lower probability sequences. It is worth noting that this speculation remains unverified at this time.

Inference Architecture

GPT-4's inference process operates on a cluster of 128 GPUs, distributed across multiple data centers in different locations. This infrastructure employs 8-way tensor parallelism and 16-way pipeline parallelism to maximize computational efficiency. Each node, comprising 8 GPUs, accommodates approximately 130 billion parameters. With a model size of 120 layers, GPT-4 can fit within 15 different nodes, possibly with fewer layers in the first node due to the need to compute embeddings. These architectural choices facilitate high-performance inference, demonstrating OpenAI's commitment to pushing the boundaries of computational efficiency.

Dataset Size and Composition

GPT-4 was trained on an impressive 13 trillion tokens, providing it with an extensive corpus of text to learn from. However, not all tokens can be accounted for by the known datasets used during training. While datasets like

Rumours and Speculations

Speculations have emerged regarding the origin of this undisclosed data. One rumor suggests that it includes content from popular platforms such as Twitter, Reddit, and YouTube, highlighting the potential influence of user-generated content in shaping GPT-4's knowledge base. Additionally, there are conjectures surrounding the inclusion of expansive collections like LibGen, a repository of millions of books, and Sci-Hub, a platform providing access to numerous scientific papers. The notion that GPT-4 was trained on the entirety of GitHub has also circulated among AI enthusiasts.

The Reporter's Opinion

Although there are many rumors, it is important to approach these rumors with caution. The training of GPT-4 may have benefited greatly from a special dataset made up of college textbooks. This dataset, which covers a wide range of courses and subjects, could have been painstakingly assembled by hand. College textbooks provide a structured and comprehensive knowledge base that can be successfully used to train a language model and are easily convertible to text files. The inclusion of such a dataset might give the impression that GPT-4 is knowledgeable in a variety of fields.

and even recall unique identifiers from platforms like Project Euler. Researchers have attempted to extract memorized sections of books from GPT-4 to gain insights into its training, further fueling curiosity about the model's inner workings. These discoveries highlight the astonishing capacity of GPT-4 to retain information and underscore the impressive capabilities of large-scale language models.

The Versatility of GPT-4

The broad spectrum of topics and fields that GPT-4 can seemingly engage with showcases its versatility. Whether it be answering complex questions in computer science or delving into philosophical debates, GPT-4's training on a diverse dataset equips it to engage with users from various domains. This versatility stems from its exposure to a vast array of textual resources, making it a valuable tool for a wide range of users.

Read more about AI:

The Evolution of Chatbots from T9-Era and GPT-1 to ChatGPT

LLaMa with 7 Billion Parameters Achieves Lightning-Fast Inference on Apple M2 Max Chip

NVIDIA Stock Surges 26% After Strong Q1 Earnings Led by Growing Demand for AI GPUs

Y Combinator's Winter 2023 Demo Day Batch Features 60 AI and Machine Learning Startups

Disclaimer

In line with the Trust Project guidelines, please note that the information provided on this page is not intended to be and should not be interpreted as legal, tax, investment, financial, or any other form of advice. It is important to only invest what you can afford to lose and to seek independent financial advice if you have any doubts. For further information, we suggest referring to the terms and conditions as well as the help and support pages provided by the issuer or advertiser. MetaversePost is committed to accurate, unbiased reporting, but market conditions are subject to change without notice.

About The Author

Damir Yalalov

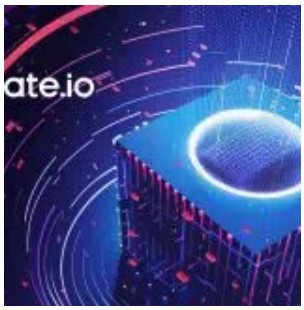
Damir is the team leader, product manager, and editor at Metaverse Post, covering topics such as AI/ML, AGI, LLMs, Metaverse, and Web3-related fields. His articles attract a massive audience of over a million users every month. He appears to be an expert with 10 years of experience in SEO and digital marketing. Damir has been mentioned in Mashable, Wired, Cointelegraph, The New Yorker,

and the CIS as a digital nomad. Damir earned a bachelor's degree in physics, which he believes has given him the critical thinking skills needed to be successful in the ever-changing landscape of the internet.

More articles

Read More

Read more



PRESS RELEASES
BUSINESS MARKETS
TECHNOLOGY

The Red Ocean of the Derivatives Market: How Gate.io Stands Out

by Victoria d'Este January 1, 2025

OPINION BUSINESS
MARKETS SOFTWARE
TECHNOLOGY

Squid Game-Inspired Tokens Flood the Crypto Market as Investors Face Growing Scam Risks

by Victoria d'Este January 1, 2025

BUSINESS MARKETS
NEWS REPORT
TECHNOLOGY

Vitalik Buterin Donates $170K in ETH to Tornado Cash Developers' Legal Fund

by Victoria d'Este January 1, 2025

BUSINESS MARKETS
NEWS REPORT
TECHNOLOGY

Jordan Approves Blockchain Policy to Modernize Government Services

by Victoria d'Este January 1, 2025

Affiliate Relationships

Bitcoin Community

bitcoin ecosystem

Bitcoin Halving

Bitcoin staking

Bitcoin trading