

[Open in app](#)

Search

[Write](#)

Introducing Retentive Networks (RetNet): A Groundbreaking Architecture for Next-Gen Language Models

Multiplatform.AI · [Follow](#)

3 min read · Jul 22, 2023



58

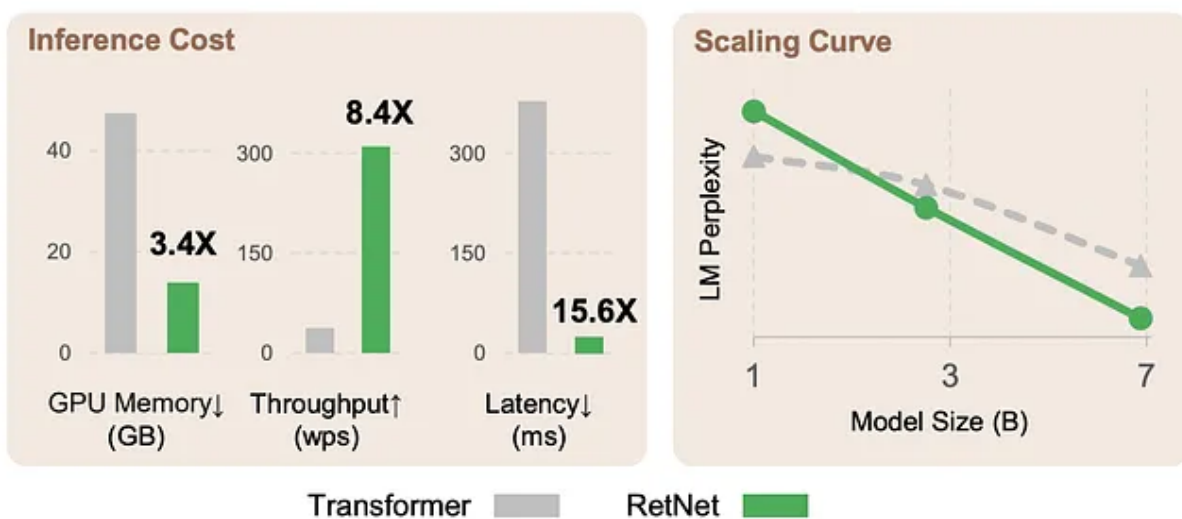


Figure 1: Retentive network (RetNet) achieves low-cost inference (i.e., GPU memory, throughput, and latency), training parallelism, and favorable scaling curves compared with Transformer. Results of inference cost are reported with 8k as input length. Figure 6 shows more results on different sequence lengths.

TL;DR:

- Retentive Networks (RetNet) introduces a revolutionary architecture for large language models.
- It achieves training parallelism, low-cost inference, and competitive performance, overcoming the limitations of traditional Transformers.
- The multi-scale retention mechanism and recurrent representations optimize GPU utilization and memory efficiency.
- RetNet excels in long-sequence modeling by cleverly encoding global and local blocks iteratively.
- Extensive trials demonstrate RetNet's impressive scaling curves and in-context learning capabilities.
- It outperforms Transformers in decoding speed, memory usage, and training acceleration.
- RetNet's inference latency remains unaffected by batch size variations, offering exceptionally high throughput.

Main AI News:

In the fast-paced world of language modeling, innovation is the key to success. Over the years, Transformer architecture has emerged as the go-to solution for building large language models. However, it comes with its own set of limitations, hindering deployment and impacting inference performance. Today, we present to you a game-changer in the field of language models — Retentive Networks (RetNet).

Addressing the shortcomings of traditional Transformers, RetNet pioneers a unique approach that achieves training parallelism, low-cost inference, and exceptional performance. The “impossible triangle” challenge, depicted in Figure 1, has long plagued researchers striving to find an optimal solution. Nonetheless, the collaboration between Microsoft Research and Tsinghua University has culminated in the birth of RetNet.

So, what makes RetNet stand out? Let’s delve into its groundbreaking features.

Training Parallelism and Efficient Inference: RetNet introduces a multi-scale retention mechanism, leveraging three processing paradigms — similar, recurrent, and chunkwise recurrent representations. By replacing the conventional multi-head attention, RetNet unlocks the true potential of GPU devices, fully utilizing their capabilities for parallel model training. This remarkable feat is achieved through efficient parallel representation and recurrent inference, leading to an $O(1)$ complexity in terms of memory and computation. The deployment cost and latency are significantly reduced, thanks to this novel approach.

Effective Long-Sequence Modeling: With the chunkwise recurrent representation, RetNet achieves effective long-sequence modeling. By encoding global blocks iteratively, RetNet conserves valuable GPU memory, while simultaneously encoding each local block to expedite processing. This ingenious technique results in remarkable performance gains.

Unmatched Performance: Extensive trials comparing RetNet with Transformers and its derivatives have validated its prowess. In language modeling experiments, RetNet consistently competes favorably, showcasing impressive scaling curves and in-context learning capabilities. Furthermore,

RetNet's inference cost remains invariant across different sequence lengths, proving its versatility.

Impressive Metrics: RetNet has proven its mettle in quantitative metrics. It boasts a staggering 8.4x faster decoding speed and utilizes a remarkable 70% less memory than Transformers equipped with key-value caches, even with a 7B model and an 8k sequence length. Not stopping there, RetNet also outperforms highly optimized FlashAttention, while saving 25–50% more memory during training acceleration. Its inference latency remains unaffected by batch size variations, delivering an unmatched throughput.

Embracing the Future: With its fascinating features, RetNet emerges as a formidable replacement for big language models. Its efficiency, scalability, and remarkable performance make it a clear choice for businesses seeking state-of-the-art language models.

Conclusion:

Retentive Networks (RetNet) brings a paradigm shift to the language modeling market. With its innovative architecture, it addresses crucial challenges faced by traditional Transformers, offering unparalleled performance and efficiency. Businesses looking to stay ahead in the competitive landscape should consider adopting RetNet as their language model solution to reap the benefits of its superior capabilities.

Source

AI

Aitechnology

Artificialintelligence

Decodingspeed

Inferencelatency



Written by Multiplatform.AI

704 Followers

Follow

One-of-a-kind AI news project dedicated to bringing you the latest breakthroughs in AI development. #AI #AINews #ArtificialIntelligence #GPT4 #AITech #OpenAI

More from Multiplatform.AI



 Multiplatform.AI

Harvard University introduced 5 free online AI courses via Coursera

3 min read · Dec 27, 2023



132



2



 Multiplatform.AI

Microsoft Unveils GPT-RAG: A Cutting-Edge Machine Learning...

TL;DR:

3 min read · Dec 20, 2023



103



2



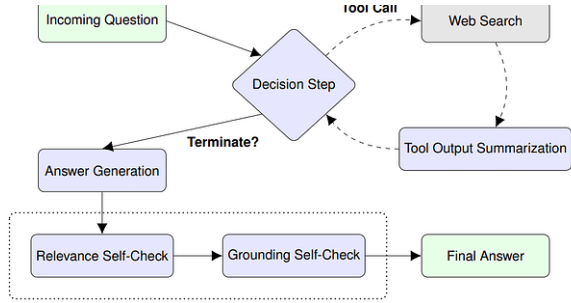


Figure 2: A state machine of the Search Agent flow. Each blue shape corresponds to a single LLM



AI

Multiplatform.AI

Introducing the Next Generation of AI: Google’s ReAct-Style LLM...

3 min read · Dec 21, 2023

57

1

AI

Multiplatform.AI

Apple’s MLX Framework: Transforming Machine Learning f...

3 min read · Dec 10, 2023

20

See all from Multiplatform.AI

Recommended from Medium



 BoredGeekSociety


Finally! 7B Parameter Model beats GPT-4!

We are entering the era of small & highly efficient models!

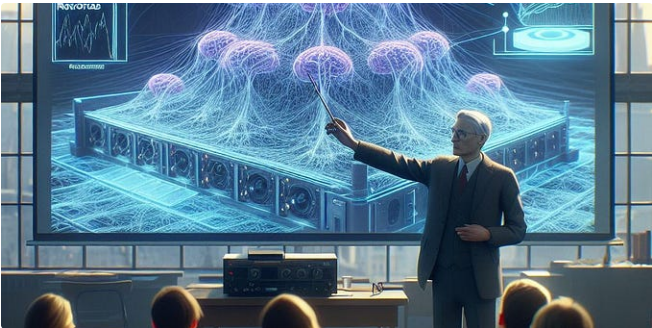
 · 2 min read · Feb 6, 2024


 689

 7







 Austin Star... in Artificial Intelligence in Plain Engli...

Reinforcement Learning is Dead. Long Live the Transformer!

Large Language Models are more powerful than you imagine

8 min read · Jan 14, 2024

 1.1K

 32





Lists



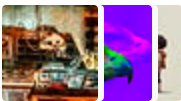
Generative AI Recommended Reading

52 stories · 736 saves



AI Regulation

6 stories · 320 saves






What is ChatGPT?

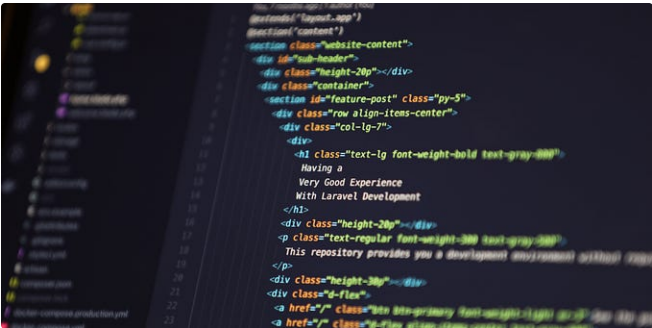
9 stories · 295 saves



ChatGPT prompts

42 stories · 1135 saves

Training Size	Compute Size	Model Size
# of Book shelves for 13T tokens	Compute time for 2.15 e25 FLOPs	Size of Excel Sheet for 1.8T params
650 kms Long line of Library Shelves	7 million years On mid-size Laptop (100GFLOPs)	30,000 Football Fields sized Excel Sheet
		
100000 tokens per Book 650 Books per shelf	100GFLOPs per second	1x1 cm per Excel cell





Rohan Balkondekar

The full training run of GPT-5 has gone live

We can expect it to be released in November, maybe on the 2nd anniversary of the...

4 min read · Jan 28, 2024



879



8



azhar in azhar labs

Building Mamba from Scratch: A Comprehensive Code Walkthrough

In the realm of deep learning, sequence modeling remains a challenging task, often...



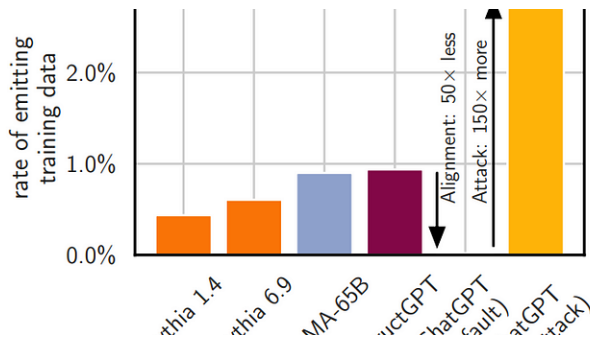
12 min read · Dec 29, 2023



428



1



Devansh in DataDrivenInvestor

Google extracted ChatGPT's Training Data using a silly trick.

Scalable Extraction of Training Data from (Production) Language Models

13 min read · Jan 8, 2024



2K



15



Cristian Leo in Towards Data Science

The Math behind Adam Optimizer

Why is Adam the most popular optimizer in Deep Learning? Let's understand it by diving...

16 min read · Jan 31, 2024



1.8K



13

[See more recommendations](#)