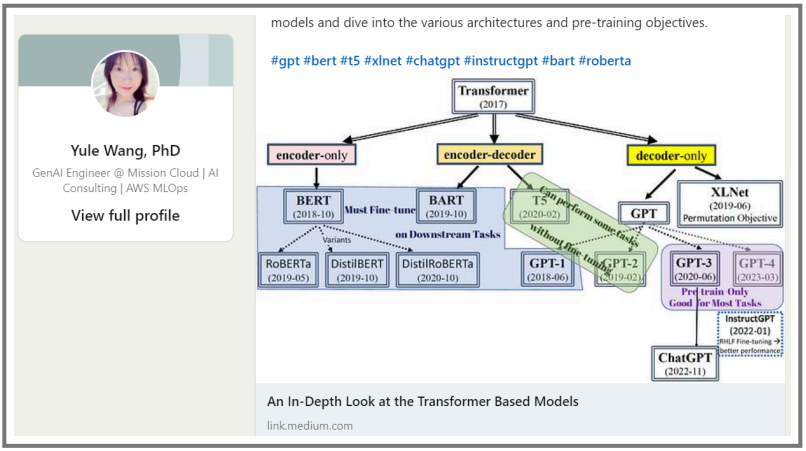
Assorted topics

Part 3 ["there's EVEN MORE!"]

Attention!!!



"A Transformer computes self-attention."

Jay Alammar: https://jalammar.github.io/illustrated-transformer/

Yule Wang: medium.com-Step-by-Step Illustrated Explanations of Transformer.pdf

Stefania: https://machinelearningmastery.com/the-transformer-attention-mechanism/

Vinija: https://vinija.ai/models/LLM/#google_vignette

Stephen Wolfram: https://writings.stephenwolfram.com/2023/02/what-is-chatgpt-doing-and-why-does-it-work/

https://www.youtube.com/watch?v=NzLwHcqE6Jw [The animated Transformer: the Transformer model explained the fun way!]

Damien Benveniste: https://learn.theaiedge.io/p/introduction-to-transformers-for-large-large-models and https://betterprogramming.pub/chatgpt-llms-and-foundation-models-a-closer-look-into-the-hype-and-implications-for-startups-b2f1d82f4d46#86bb

Transformers can be adapted to other domains, eg. ViT [https://www.linkedin.com/pulse/vision-transformer-damien-benveniste-phd-nygoc], time-series prediction, music gen, etc.

Context

Because 'context is everything', we need to address the 'quadratic bottleneck' of attention computation. Here are ways:

> Indeed, in DNA models like EVO, the researchers used the Hyena operator instead of attention to avoid the previously mentioned quadratic relationship. Hyena operators use **long convolutions** instead of attention to capture long-range dependencies with a subquadratic cost.

> *and cheaper.*
>
> Other alternatives aim for a hybrid approach that, instead of ditching attention altogether, finds the sweet spot between attention and other operators to maintain performance with lower costs.
>
> Recent examples include Jamba, which cleverly mixes Transformers with other, more efficient architectures like *Mamba*.
>
> > *Mamba, Hyena, Attention… you are probably thinking I'm just name-dropping fancy words to prove a point.*

AlphaSignal

Hey ,

Welcome to today's edition of AlphaSignal.

Whether you are a researcher, engineer, developer, or data scientist, our summaries ensure you're always up-to-date with the latest breakthroughs in AI.

Let's get into it!

Lior

IN TODAY'S SIGNAL

- **Top Paper:** Google New Infinite Context Method
- **Trending Repos:**
  - gemini-cookbook
  - aider
  - ragflow
- **Top Lecture:** Pretraining an LLM on Unlabeled Data

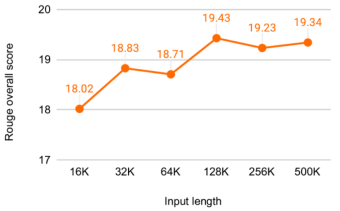
Read Time: 4 min 06 sec

TOP NEWS

Language Models

Google Releases New Infinite Context Method

⇧ 1683 ⇆ 394



What's New

Can LLMs Handle Unlimited Context? Google researchers introduced a new concept called Infini-attention in their latest paper, enabling LLMs to process inputs of any length.

Comparison with Traditional Transformers: Typical transformers reset their attention memory after each context window to manage new data, losing previous context. For example, in a 500K token document split into 100K token windows, each segment starts fresh without memory from the others.

Infini-attention's Approach: Infini-attention retains and compresses the attention memory from all previous segments. This means in the same 500K document, each 100K window maintains access to the full document's context.

The model compresses and reuses key-value states across all segments, allowing it to pull relevant information from any part of the document.

How Infini-attention Works:

- Utilizes standard local attention mechanisms found in transformers.
- Integrates a global attention mechanism through a compression technique.
- Merges both local and global attention to manage extended contexts efficiently.

In other words, the method effectively gives each window a view of the entire document, achieving what's termed as "infinite context."

Key Performance Metrics:

- **1B Model:** Effectively manages sequences up to 1 million tokens.
- **8B Model:** Achieves state-of-the-art results in tasks like summarizing books up to 500K tokens in length.

Key Highlights:

- **Memory Efficiency:** Constant memory footprint regardless of sequence length.
- **Computational Efficiency:** Reduces computational overhead compared to standard mechanisms.
- **Scalability:** Adapts to very long sequences without the need for retraining from scratch.

Why This Matters

Elvis Saravia: *"Given how important long-context LLMs are becoming having an effective memory system could unlock powerful reasoning, planning, continual adaption, and capabilities not seen before in LLMs. Great paper!"*

Community Feedback

swyx: *"with Griffin and Infini-attention, it increasingly feels like Google leapfrogged Together and RWKV in the race for scaling up linear attention, and they shared a watercooler conversation with Anthropic or something"*

Raúl Avilés Poblador: *"Sounds good on paper but scaling this to near-infinity would also mean inference would require crazy hardware resources, isn't it?"*

READ THE PAPER

Get the Pytorch Implementation

TOP OF GITHUB

Language Models

gemini-cookbook

A collection of guides and examples for the Gemini API, including quickstart tutorials for writing prompts and using different features of the API, and examples of things you can build.

☆ 1573

Code Assistants

aider

Aider is a command line tool that lets you pair program with GPT-3.5/GPT-4, to edit code stored in your local git repository. Aider will directly edit the code in your local source files, and git commit the changes with sensible commit messages. You can start a new project or work with an existing git repo. Aider is unique in that it lets you ask for changes to pre-existing, larger codebases.

☆ 8922

RAG

ragflow

RAGFlow is an open-source RAG (Retrieval-Augmented Generation) engine based on deep document understanding. It offers a streamlined RAG workflow for businesses of any scale, combining LLM (Large Language

Models) to provide truthful question-answering capabilities, backed by well-founded citations from various complex formatted data.

☆ 4709

Give Your Honest Opinions on Building with AI

How satisfied are you with your current AI stack? Do you think your company is allocating GPUs correctly? What are you most excited to see in the next generation of models and tools?

Share your thoughts on the evolving state of AI in the newest State of AI survey. (You could win a $500 Amazon gift card!)

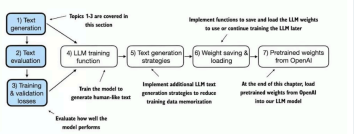
Share thoughts ↗

TOP LECTURE

Language Models

Build an LLM from Scratch Chapter 5: "Pretraining an LLM on Unlabeled Data"

☆ 13,532



What's New

Chapter 5 of Sebastian Raschka's "Build an LLM from Scratch" book, titled "Pretraining an LLM on Unlabeled Data," is now available. This chapter advances the series by focusing on the implementation of a training function and the initiation of pretraining for the LLM.

Key topics covered include:

- Computing the training and validation set losses to assess the quality of text generated by the LLM during training.
- Implementing a training function and starting the pretraining process.
- Techniques for saving and loading model weights, allowing for the continuation of training at different stages.
- Loading pretrained weights from OpenAI to enhance model performance.

CHECK THE REPO

How was today's email?

Not Great Good Amazing

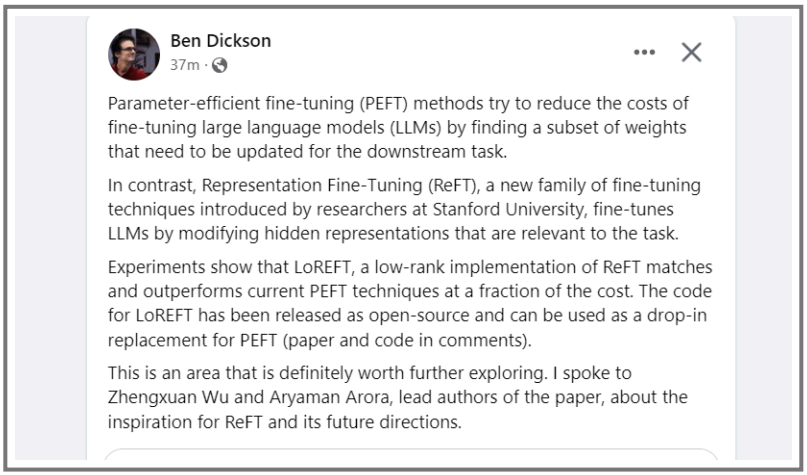
Ghita is an editor at AlphaSignal and AI Engineer at Clear Ventures. Experienced in Computer Vision and Language models, she holds 2 masters in Applied Mathematics.

Thank You.

PFFFFF-FT!!

A way to augment an LLM's general learning, is to 'fine tune' (FT) it for specialized domains such as medicine, law, etc.

ReFT is the latest FT technique:



RAGGGG-GG

Another way to augment a query's context is to use external memory - this can be KG, DBs, text...

But there can be 'issues' - context mismatch, incorrectly chunked input...

genAI

GANs start out as a pair of 'adversarial' networks - the 'student' one gets VERY good at **content generation**!!!

VAEs do likewise.

GPT: **Generative** [token generation] Pretrained Transformer.

Today we have variations - generation via diffusion, generation via radiance modeling, generation via point-cloud modeling, generation via musical pattern modeling...

Almost ANYTHING can be generated - molecular representations, architectural plans, circuit layout, business cards, A/V/I/3D, etc etc.

New directions

As is to be expected, the wild new world of LLMs began with 'ChatGPT' in Nov'22 - but is expanding in multiple, very interesting directions!

- hardware
- MCC
- FT alts
- RAG2.0, no-RAG...
- SLMs
- MoE
- bit-crushed weights
- multimodal LLMs
- 'embodied' LLMs
- agents
- ...

Issues

The PRACTICE of 'searching' for (or looking up) info has become part of how we live.

Given that, companies such as Google that enable this for us, have enormous POWER over the process (ie. how they serve us results).

Search results can be 'manipulated' (ie. altered) before being sent to us.

Also, searching might bump up against legal protections afforded to content.

On top of it all, the use of LLMs to serve (summarize) search results introduces one more wrinkle: how would we know what came from the LLM itself, and what came from search? This insightful report details how LLMs (for example) help companies concentrate power.