GPT (to be changed)

Michael Gathara
University of Alabama at Birmingham

Akshar Patel
University of Alabama at Birmingham

Vaishak Menon
University of Alabama at Birmingham

Jason Liu
University of Alabama at Birmingham

Elisabeth Molen
University of Alabama at Birmingham

Trenton Davis
University of Alabama at Birmingham

***Abstract*—Large language models (LLMs) are a subset of artificial intelligence (AI) models that have the ability to intake, understand, and generate human readable text. Since their inception, these models have found rapid success and widespread adoption in a vast range of industries and use cases including software engineering, healthcare, research, and supply-chain management. A specific type of LLM- generative pretrained transformers (GPTs)- stand out as one of the most used and adopted models as they are tuned to be highly effective at understanding conversational context and generating coherent and relevant text. Built on top of the transformer architecture, GPTs are often more finely tuned to perform specific tasks after first being trained on a large corupus of less specific textual data- hence the pre-trained portion of the name. While the transformer has remained the core of these models, many different optimizations have been proposed and implemented over the past several years including flash attention and multi-layer attention (MLA). In this work, we implement a scratch-built GPT equipped with some of these optimizations, train it on the fineweb dataset, and then test it on several tasks and analyze some of the features of the model.**

I. Introduction

In recent years, artificial intelligence (AI) models have been gaining popularity in both personal use-cases as well as in many different industries. Among these, none seem to have risen to such a high level of popular adoption as large language models (LLMs) which are AI models that can intake, understand, and generate human-readable text [1]. Usage of these has gained traction in a broad range of industries including research, healthcare, supply-chain, and software engineering [2] [3] [4]. The most powerful LLMs are often built on an architectural framework called a transformer which calculates importance of different parts of the input and weighs the relationship between them in order to produce more coherent and relevant outputs. In their seminal 2017 paper, Vaswani et al. introduced the transformer architecture and set the stage for the many iterations of LLMs that have come in the years since [5]. This original paper focused on the core attention component of the transformer which calculates the interactions between different input tokens and uses this contextual knowledge to more dynamically generate relevant output text. While a transformative architecture, the core attention mechanism does suffer from a few key drawbacks that have motivated the need for some important optimizations in the years since. Because it calculates interactions between every input token pair, the attention mechanism suffers from a quadratic time and memory complexity leading to some amnesia inducing context length limitations. Additionally, models immediately following the transformer architecture follow a single uniform application of the attention mechanism at their different layers. This lack of heterogenity and depth didn't allow the models to fully capture the nuanced and dynamic interactions between the input tokens.

Generative pre-trained transformers (GPTs) are a specific instantiation of LLMs that use the transformer architecture as a backbone. GPTs are engineered to be highly effective at generating relevant text and at performing conversationally oriented tasks by first pre-training them on a huge corpus of text and then fine tuning them for their specific tasks- a process that was introduced by Radford et al. [6]. Following that initial introduction, GPTs utilizing magnitudes more parameters began to be explored- consistently increasing their usefulness in conversational tasks. Specifically, post-training capacity of the models to learn and respond to instructions deliverd to them as inputs was detailed by Brown et al. in their 2020 paper introducing GPT-3 [1]. While making significant bounds in model performance and practical usefulness, these architectures still suffered from the same core limitations of the original transformer setup that they are built on top of.

In the present work, we implement a from-scratch trained GPT architecture but we also explore some of the optimizations that have been introduced to address the aformentioned limitations of the transformer architecture. Two of the core optimizations that we utilize are flash-attention and multi-layer attention. Flash attention addresses the issue of quadratic memory complexity in the original attention mechanism- an issue that significantly slows down the computation as expensive GPU memory transfers abound with full attention scores being computed and scored. Flash attention focuses on making the attention mechanism input/output (IO) aware and splits the computations into tiles that can fit on a GPUs on-chip memory and eschew temporally costly IO operations [7]. In our setup, we use flash attention and see significant improvements in our model training time as well as inference

II. Methods

A. Model Architecture

B. Training

C. Evaluation and Introspection

III. Results

IV. Discussion

V. Conclusion

References

[1] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, "Language models are few-shot learners," 2020. [Online]. Available: https://arxiv.org/abs/2005.14165

[2] W. Liang, Y. Zhang, Z. Wu, H. Lepp, W. Ji, X. Zhao, H. Cao, S. Liu, S. He, Z. Huang, D. Yang, C. Potts, C. D. Manning, and J. Y. Zou, "Mapping the increasing use of llms in scientific papers," 2024. [Online]. Available: https://arxiv.org/abs/2404.01268

[3] C. Zhang, Q. Xu, Y. Yu, G. Zhou, K. Zeng, F. Chang, and K. Ding, "A survey on potentials, pathways and challenges of large language models in new-generation intelligent manufacturing," *Robotics and Computer-Integrated Manufacturing*, vol. 92, p. 102883, 2025. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0736584524001704

[4] A. Urlana, C. V. Kumar, A. K. Singh, B. M. Garlapati, S. R. Chalamala, and R. Mishra, "Llms with industrial lens: Deciphering the challenges and prospects – a survey," 2024. [Online]. Available: https://arxiv.org/abs/2402.14558

[5] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," 2023. [Online]. Available: https://arxiv.org/abs/1706.03762

[6] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving language understanding by generative pre-training," 2018.

[7] T. Dao, D. Y. Fu, S. Ermon, A. Rudra, and C. Ré, "Flashattention: Fast and memory-efficient exact attention with io-awareness," 2022. [Online]. Available: https://arxiv.org/abs/2205.14135