Optimizing Large Language Models To Minimize Attention Latency For Increasingly Complex Inputs

Kiera Lai, Computer Systems Engineering
Mentor: Ryan Meuth Ph.D., Associate Teaching Professor
School of Computing and Augmented Intelligence

Research Question:

How can the context window of large language models be modified to improve the attention lengths of artificial intelligence systems?

Background:

Large Language Models are foundational in Generative Artificial Intelligence systems. They are mainly used to train the systems to analyze, summarize, predict, and produce human-like text. One of the issues with artificial intelligence is being able to make coherent and relevant statements given lengthier conversational inputs.

Abstract:

This project aims to improve Large Language Models' context retention and coherence by optimizing attention mechanisms, altering parameters, and amending inputs.

Simplified generative model

Develop word relation graph

Adjust parameters and functions

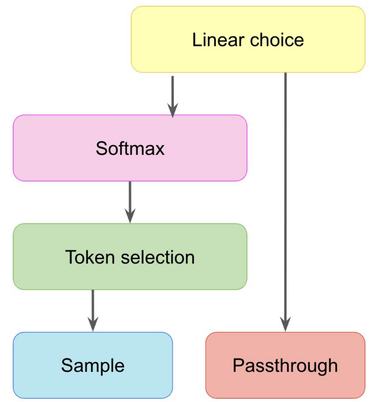
Input heavy inputs

$$W_{\text{out}} \in \mathbb{R}^{V \times d},\ b_{\text{out}} \in \mathbb{R}^{V}$$

$$\text{logits} = W_{\text{out}} \cdot h_n + b_{\text{out}}$$

$$P(\text{token}_i) = \frac{e^{\text{logits}_i}}{\sum_j e^{\text{logits}_j}}$$

$$E \in \mathbb{R}^{V \times d}$$



Analysis and process:

Optimizing flow through linearization. The output matrix connects the states of the vocabulary space. Vocabulary size generated with a random text code and inputted into a word context. This is used to find the next possible outcome. Process like a translator and a logically response is produced.

Conclusion:

Through the adjustment of the model's parameters, there are hidden states found in the vocabulary space. Efficiency of attention retention can be improved by editing input structuring. By optimizing the flow between states, the model better manages complex and lengthier inputs. The study demonstrated with simplified models and refining transformations.

Possible future work:

- Translate the model into a larger model
- Explore greater and detailed vocabularies
-

References:

[1] Agarwal, S. , Acun, B. , Hosmer, B. , Elhoushi, M. , Lee, Y. , Venkataraman, S. , Papailiopoulos D. , & Wu C.J. (2024) Chai: Clustered head attention for efficient LLM Inference, arXiv.org.

[2] Kwon, W. , Li, Z. , Zhuang, S. , Sheng, Y. , Zheng, L. , Yu, C.H. , Gonzalez, J.E. , Zhang, H,. , & Stoica, I. (2023) Efficient memory management for large language model serving with paged attention: Proceedings of the 29th Symposium on Operating Systems principles, ACM Conferences.

[3] Li, T. , Zhang, G. , Do, Q.D. , Yue, X. , & Chen, W. (2024) Long-context LLMS struggle with long in-context learning, arXiv.org.

[4] Vaswani, A. , Shazeer, N. , Parmar, N. , Uszkoreit, J. , Jones, L. , Gomez, A.N. , Kaiser, L. , & Polosukhin, I. (2023) Attention is all you need, arXiv.org.

[5] Wang, X. , Salmani, M. , Omidi, P. , Ren, X. , Rezagholizadeh, M. , & Eshaghi, A. (2024) Beyond the limits: A survey of techniques to extend the context length in large language models, arXiv.org.

Acknowledgements:

I would like to thank Professor Meuth and Professor Malpe for their guidance and support. Special thanks to Reeyan for his inspiration

FURI

Ira A. Fulton Schools of Engineering
Arizona State University