Pushing the limits of the compressive memory Introduced in Infini-attention

Architectural Decisions for Language Modelling with (Small) Transformers

Author
Lauri Kesküll

Affiliation
EEMCS, TU Delft

Supervisors
Prof. Maliheh Izadi
Prof. Arie van Deursen
Aral De Moor

References
[1] T. Munkhdalai, M. Faruqui, and S. Gopal, "Leave No Context Behind: Efficient Infinite Context Transformers with Infini-attention," arXiv preprint arXiv:2404.07143, 2024.
[2] Alammar, J. (2018). The Illustrated Transformer. Blog post. Retrieved from https://jalammar.github.io/illustrated-transformer/
[3] Poster format by Jan van der Meulen

01 Introduction

- Transformers are revolutionizing language processing
- Traditional transformers struggle with long-form content like books due to quadratic scaling
- Infini-Attention introduces a compressive memory to handle long contexts with linear computational demands.
- Exploring the viability of small models tailored for local devices is crucial for privacy, reduced latency, and efficient data usage, especially given the scarcity of new training data.
- We evaluated how to optimally integrate Infini-Attention into transformer models

02 Contribution

- Concluded whether Infini-Attention should be incorporated during pre-training or fine-tuning.
- Provided insights into the integration process of Infini-attention by analyzing the convergence behavior of gating parameters during fine-tuning.
- Created a replication package of Infini-attention for reproducing our findings.
- The replication package includes models published on HuggingFace and modifications to the Transformers library that are neccessary to support Infini-attention

03 Approach

- Infini-attention (Figure 1) combines self-attention and long-term linear attention within a single transformer block.
- It stores information from previous segments in a compressive memory

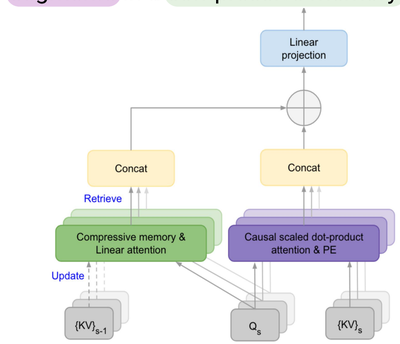

Figure 1. The Infini-attention architecture, [1]

- Self-attention computes pairwise interactions between all elements in its input
- During this process, each token evaluates the degree of attention it should allocate to other tokens (Figure 2).
- This allows the model to capture relationships between different parts of the input sequence.

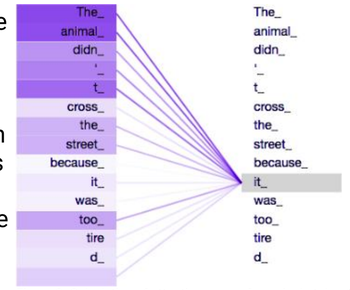

Figure 2. Calculation of attention scores for a single token [2]

- The compressive memory works thanks to associative binding
- It can keep storing information without taking up any extra space
- It can be visualized with simple XOR operations seen in Figure 3

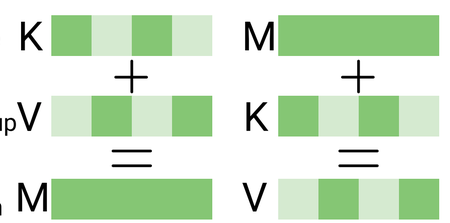

Figure 3. Example of a Key and Value being stored and retrieved from Memory

04 Experiment Results

- Fully pre-trained models with Infini-Attention slightly outperform fine-tuned models.
- Beta values determine how much the model looks at compressive memory, they seem to converge after around 0.5 epochs of fine-tuning (Figure 4).
- Models that were trained on shorter context lengths displayed higher beta values (Figure 5)
- A segment length of 128 displayed the most balanced distribution of beta values (Figure 6).

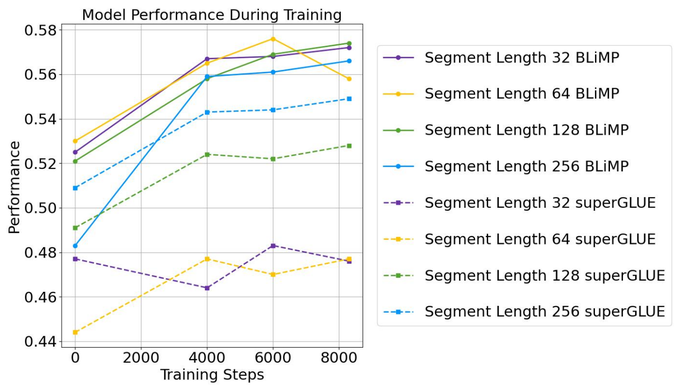

Figure 4. Evaluation scores of Infini-attention enabled GPT-Neo models throughout the fine-tuning process

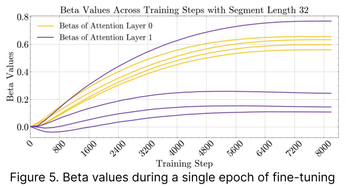

Figure 5. Beta values during a single epoch of fine-tuning

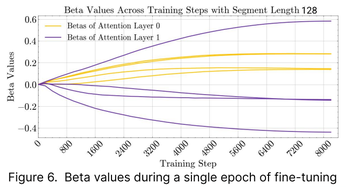

Figure 6. Beta values during a single epoch of fine-tuning

05 Discussion

- Fully pre-trained models with Infini-Attention hold a slight performance edge over fine-tuned models, likely due to consistent training conditions.
- Fine-tuning beyond 0.5 epochs likely offers diminishing returns.
- This relative meaning could simplify learning grammatical structure.
- Infini-attention may be most efficient when boosting the maximum input sequence by a factor of 4.
- Infini-Attention may be best used for 'open-book' questions, where the task is posed before the input sequence.

06 Conclusion

- The goal was to investigate the effectiveness of Infini-Attention in transformer models and find the optimal integration strategy.
- Models trained with consistent segment lengths throughout training performed best.
- Shorter context still lead to decreased performance, indicating that Infini-Attention does not fully compensate for reduced context lengths.
- Future research should focus on models pre-trained and fine-tuned with consistent segment lengths and develop evaluation metrics tailored to longer sequences to better assess Infini-Attention's potential.