Position Interpolation Improves ALiBi Extrapolation

Faisal Al-Khateeb, Nolan Dey, Daria Soboleva, Joel Hestness

Cerebras Systems   *faisal.alkhateeb@cerebras.net*

Abstract

Linear position interpolation helps pre-trained models using rotary position embeddings (RoPE) to extrapolate to longer sequence lengths. We propose using linear position interpolation to extend the extrapolation range of models using Attention with Linear Biases (ALiBi). We find position interpolation significantly improves extrapolation capability on upstream language modelling and downstream summarization and retrieval tasks.

1 Introduction

LLMs support for long sequence lengths is critical for many downstream applications. However, retraining or fine-tuning models to support long sequence lengths is costly. Recent works by Dosovitskiy et al. (2021) and Chen et al. (2023) show approaches to interpolate learned and rotary position embeddings (RoPE), respectively. These techniques improve models' capability to interpolate and extrapolate to different sequence lengths. Similarly, Attention with Linear Biases (ALiBi) (Press et al., 2021) adds a recency bias in the attention mechanism to help the model extrapolate to longer sequence lengths. However, recent work by Dey* et al. (2023) shows ALiBi position embeddings only extrapolate well to ∼12% beyond the trained sequence length for an over-trained model (Hoffmann et al., 2022). Thus, we propose to extend the sequence extrapolation capabilities of ALiBi position embeddings using position interpolation. Position interpolation with ALiBi extends sequence lengths of models up to 2x the maximum training sequence length while maintaining its original language modelling performance.

2 Position Interpolation

Linear position interpolation (PI) was proposed concurrently by Chen et al. (2023) and kaiokendev (2023) to extend the effective context length of models using rotary position embeddings (RoPE) (Su et al., 2022). In this work, we extend position interpolation to also improve the extrapolation capability of models with ALiBi position embeddings, without performing additional pre-training or fine-tuning.

2.1 ALiBi with Linear Position Interpolation

In its original implementation, ALiBi adds a bias vector to the query-key dot product. The bias vector is formulated based on predefined slopes for each head, and the positional differences between the queries and the keys. Thus, we can formulate the attention scores for head_j and query_i as:

$$\text{softmax}\left(q_i K^T + \underbrace{m_j \cdot [-(i-1), \ldots, -2, -1, 0]}_{\text{positional bias}}\right)$$

To enable ALiBi extrapolation during inference, we propose scaling the slopes dynamically by a factor of L/L' where L is the maximum sequence length observed during training and L' is the extended input sequence length during inference, $m'_j = m_j \left(\frac{L}{L'}\right)$. Noting, we only scale the slopes when $L' > L$ to maintain the

previous performance of the model for samples with sequence length smaller than or equal to the training sequence length.

In models using RoPE, attention score magnitudes tend to blow up during extrapolation (Chen et al., 2023). Hence, the motivation for using position interpolation with RoPE was to scale down positional distances to the stable and bounded interpolation range seen during training. Contrary to RoPE, we observe ALiBi introduces lower magnitude attention scores for tokens in the extrapolation regime than are seen in the interpolation regime (Figure 1). By applying position interpolation, we adjust the ALiBi slope to scale up attention scores and prevent the introduction of lower magnitudes for positional differences beyond the training context length.
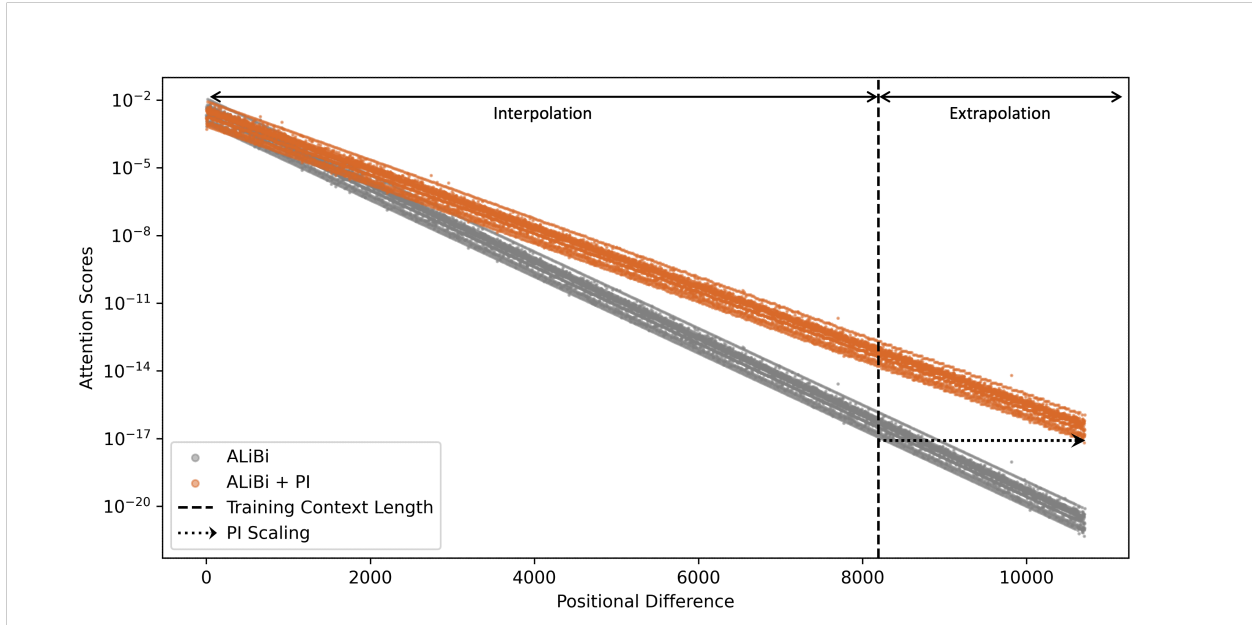

Figure 1: Attention scores in an exemplar attention head (last head of first layer) from BTLM-3B-8K model for a query at position 10.5K.

3 Evaluation

We tested ALiBi with position interpolation using BTLM-3B (Dey* et al., 2023) and MPT-7B (Team, 2023a;b) pre-trained models without fine-tuning.

3.1 Language Modeling

To quantify the improvement in extrapolation capability from using position interpolation with ALiBi position embeddings, we follow Peng et al. (2023) and measure the average perplexity on 10 documents from Proof-pile (Azerbayev et al., 2022) truncated to 16K tokens.

In Figure 2, we evaluate BTLM-3B-8K and MPT-7B-8K models and observe the baseline models can only extrapolate to 9K-10K context lengths. On the other hand, when using position interpolation these models are able to maintain the same low perplexity for up to at least 16K tokens (2x the training maximum sequence length).

Similarly, in Figure 3 we evaluate BTLM-3B-2K[1] and MPT-7B-2K and find the baseline models extrapolate to ∼2.5K-3K tokens. Position interpolation extends the extrapolation to ∼4K tokens before perplexity

---

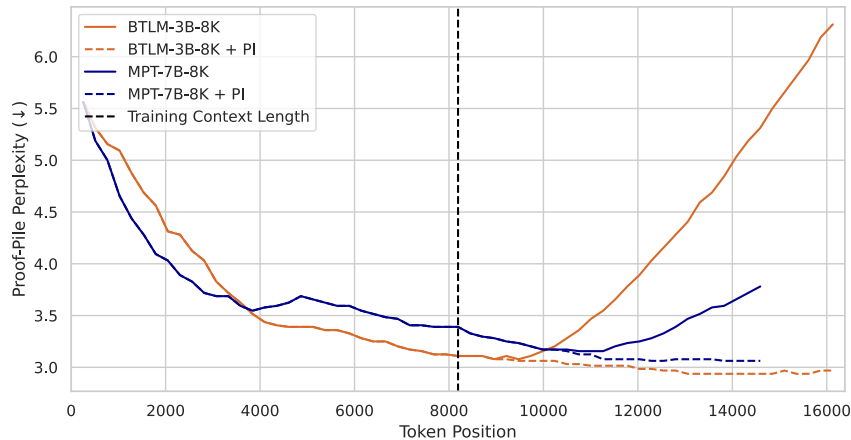[1]The final checkpoint from the 2048 context length training phase.

2

Figure 2: Average perplexity per token position over a set of ten Proof-pile documents for pre-trained models with 8K training context length.

degradation. It should also be noted that perplexity increases more slowly as token position is increased beyond the training context length.
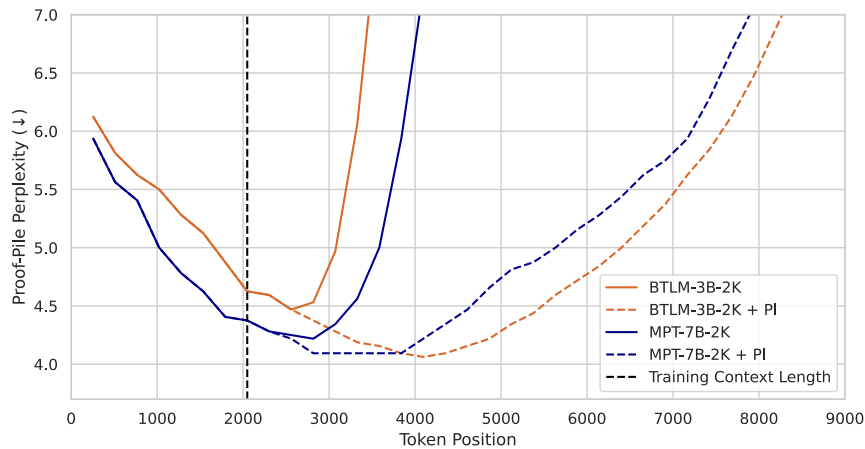

Figure 3: Average perplexity per token position over a set of ten Proof-pile documents for pre-trained models with 2K training context length.

3.2 Document Summarization Tasks

Document summarization is an important application for LLMs that requires inference at long context lengths. We evaluated BTLM-3B-8K up to 16K context lengths on GovReports (Huang et al., 2021) and QMSum (Zhong et al., 2021) summarization tasks. The two tasks require the model to understand long context documents based on samples of government reports and meeting transcripts, respectively. Table 1 shows that position interpolation significantly improves BTLM-3B-8K's ability to summarize 16K context lengths, roughly doubling ROUGE scores across tasks.

3

Model	QMSum (↑)			GovReports (↑)		
	R-1	R-2	R-L	R-1	R-2	R-L
BTLM-3B-8K	7.3	2.1	5.8	8.1	3.5	5.7
BTLM-3B-8K + PI	**16.6**	**4.7**	**12.8**	**14.7**	**7.0**	**9.9**

Table 1: ROUGE scores on the QMSum and GovReports long text summarization tasks. We only evaluate samples less than 16,384 tokens in length.

3.3 Long Range Retrieval Tasks

To evaluate long range retrieval capabilities, we use both LongEval tasks (Li* et al., 2023). The "Coarse-grained Topic Retrieval" task requires models to retrieve the first discussed topic from a long conversation that spans multiple topics, and the "Fine-grained Line Retrieval" task which requires models to precisely retrieve a number from a long document. Figure 4 shows that position interpolation greatly improves BTLM-3B-8K's ability to retrieve information from documents longer than seen during training, in both tasks. The line retrieval task being more challenging still suffered from a drop in performance when extending the document length.
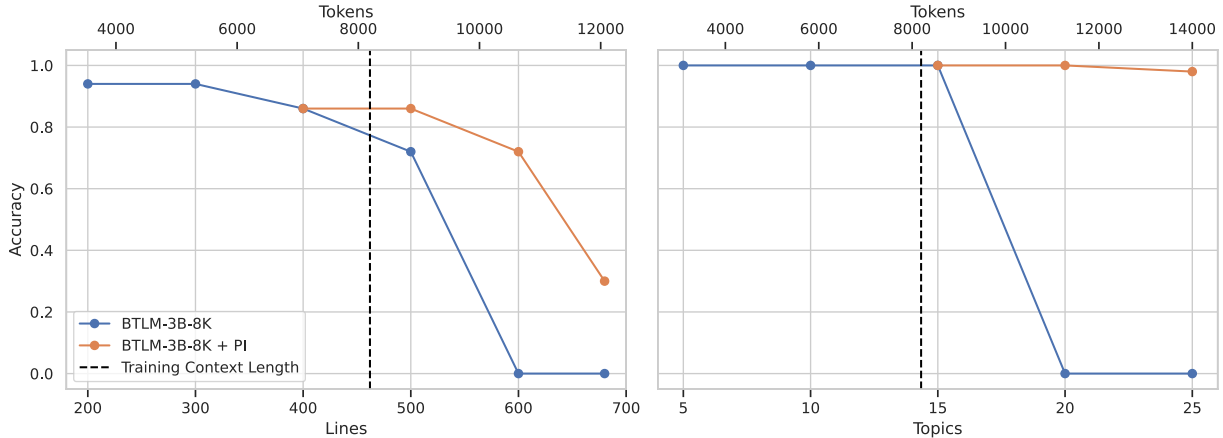

Figure 4: Accuracy on the LongEval line retrieval and topic retrieval tasks.

4 Conclusion

In this work, we extend the position interpolation method to significantly improve the extrapolation capability of models with ALiBi position embeddings. With no additional training, we demonstrate this improvement on long context language modeling, document summarization, and retrieval tasks. As future work, this combination of ALiBi position embeddings and position interpolation could be used to achieve further improvements by fine-tuning models with longer context lengths Chen et al. (2023) than seen during training.

References

Zhangir Azerbayev, Edward Ayers, and Bartosz Piotrowski. Proof-pile, 2022. URL https://github.com/zhangir-azerbayev/proof-pile.

Shouyuan Chen, Sherman Wong, Liangjian Chen, and Yuandong Tian. Extending Context Window of Large Language Models via Positional Interpolation, 2023. URL https://arxiv.org/abs/2306.15595.

Nolan Dey*, Daria Soboleva*, Faisal Al-Khateeb, Bowen Yang, Ribhu Pathria, Hemant Khachane, Shaheer Muhammad, Zhiming (Charles) Chen, Robert Myers, Jacob Robert Steeves, et al. BTLM-3B-8K: 7B Parameter Performance in a 3B Parameter Model, 2023. URL https://arxiv.org/abs/2309.11568.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, 2021. URL https://arxiv.org/abs/2010.11929.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. An Empirical Analysis of Compute-optimal Large Language Model Training. In *The Conference on Neural Information Processing Systems (NeurIPS)*, 2022.

Luyang Huang, Shuyang Cao, Nikolaus Parulian, Heng Ji, and Lu Wang. Efficient Attentions for Long Document Summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2021. URL https://aclanthology.org/2021.naacl-main.112.

kaiokendev. Things I'm learning while training superhot, 2023. URL https://kaiokendev.github.io/til#extending-context-to-8k.

Dacheng Li*, Rulin Shao*, Anze Xie, Ying Sheng, Lianmin Zheng, Joseph E. Gonzalez, Ion Stoica, Xuezhe Ma, and Hao Zhang. How Long Can Open-Source LLMs Truly Promise on Context Length?, 2023. URL https://lmsys.org/blog/2023-06-29-longchat.

Bowen Peng, Jeffrey Quesnelle, Honglu Fan, and Enrico Shippole. YaRN: Efficient Context Window Extension of Large Language Models, 2023. URL https://arxiv.org/abs/2309.00071.

Ofir Press, Noah A. Smith, and Mike Lewis. Train Short, Test Long: Attention with Linear Biases Enables Input Length Extrapolation, 2021. URL https://arxiv.org/abs/2108.12409.

Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. RoFormer: Enhanced Transformer with Rotary Position Embedding, 2022. URL https://arxiv.org/abs/2104.09864.

MosaicML NLP Team. Announcing MPT-7B-8K: 8K Context Length for Document Understanding, 2023a. URL https://www.mosaicml.com/blog/long-context-mpt-7b-8k.

MosaicML NLP Team. Introducing mpt-7b: A new standard for open-source, commercially usable llms, 2023b. URL www.mosaicml.com/blog/mpt-7b.

Ming Zhong, Da Yin, Tao Yu, Ahmad Zaidi, Mutethia Mutuma, Rahul Jha, Ahmed Hassan Awadallah, Asli Celikyilmaz, Yang Liu, Xipeng Qiu, and Dragomir Radev. QMSum: A New Benchmark for Query-based Multi-domain Meeting Summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2021. URL https://aclanthology.org/2021.naacl-main.472.