


Zero-Training Context Extension for Transformer Encoders via Nonlinear Absolute Positional Embeddings Interpolation

Ivan Danylenko 

kowd.pauuh@gmail.com

Abstract. Our code and models are publicly available¹.

Keywords: Transformer Encoders · Context Extension · Absolute Positional Embeddings · Nonlinear Interpolation.

1 Introduction

Transformer encoders are typically trained with a fixed maximum sequence length (context). The most common choice is a limit of 512 tokens [2,4], beyond which models cannot directly process text. This is a bottleneck for applications requiring long documents processing, where relevant information may appear well beyond the context window of the model. Extending the context length through retraining or fine-tuning has proven either ineffective or computationally expensive [5].

To address this limitation, recent research explores positional embedding interpolation techniques, which map longer sequences into the length range the model was trained on. Position Interpolation (PI) proposed by Chen et al. [1] and further developed YaRN method by Peng et al. [3] allowed for successful context extension of the models trained with Rotary Position Embeddings (RoPE), though these often require fine-tuning. Relatively fewer studies explore context extension of the models which leverage Absolute Positional Embeddings (APE). Zhu et al. inspired by Position Interpolation [1] applied its variant – Linear Position Interpolation [6] – to models with APE, yielding improvement on *LongEmbed* [6] benchmark after fine-tuning interpolated embeddings.

In this work, we propose a training-free method for extending the context length of transformer encoders with absolute positional embeddings. We show that positional embedding vectors in APE-based models follow nonlinear patterns across dimensions, which linear interpolation fails to capture. Motivated by this, we apply per-dimension interpolation of the pretrained embedding matrix using cubic splines. Unlike linear interpolation, which imposes a fixed scaling across position indices, our approach allows for arbitrary context lengths without any model retraining or adaptation.

¹ <https://github.com/Kowd-PauUh/encoders-context-extension>

We empirically evaluate the robustness of our method on *LongEmbed* benchmark and achieve average $XXXNdcg@10$ p.p. performance improvement over Linear Position Interpolation. Beyond this, we investigate how far the context can be extended without performance degradation, and analyze the conditions under which degradation begins to occur. Our results offer a deeper understanding of the geometry of positional embeddings and their role in long-sequence generalization.

2 Related Work

3 Proposed Method

4 Results

5 Discussion

References

1. Chen, S., Wong, S., Chen, L., Tian, Y.: Extending context window of large language models via positional interpolation. arXiv preprint arXiv:2306.15595 (2023)
2. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers). pp. 4171–4186 (2019)
3. Peng, B., Quesnelle, J., Fan, H., Shippole, E.: Yarn: Efficient context window extension of large language models. arXiv preprint arXiv:2309.00071 (2023)
4. Wang, L., Yang, N., Huang, X., Jiao, B., Yang, L., Jiang, D., Majumder, R., Wei, F.: Text embeddings by weakly-supervised contrastive pre-training. arXiv preprint arXiv:2212.03533 (2022)
5. Zhao, L., Feng, X., Feng, X., Zhong, W., Xu, D., Yang, Q., Liu, H., Qin, B., Liu, T.: Length extrapolation of transformers: A survey from the perspective of positional encoding. arXiv preprint arXiv:2312.17044 (2023)
6. Zhu, D., Wang, L., Yang, N., Song, Y., Wu, W., Wei, F., Li, S.: Longembed: Extending embedding models for long context retrieval. arXiv preprint arXiv:2404.12096 (2024)