A Comparative Study of RoPE-based Positional Encodings from A Scaling Perspective

Zhu Zhang
Tsinghua University

Tianxing Yang
Tsinghua University

Zihan Yan
Tsinghua University

1 Background and Motivation

Transformers [11], as the backbone of Large Language Models (LLMs), have become the dominant architecture for natural language processing tasks. However, their quadratic computational complexity makes training on long sequences inefficient and resource-intensive. A common solution involves pre-training the model on shorter sequences (e.g., 4k tokens) to develop initial capabilities, followed by further pre-training on longer sequences (e.g., 32k tokens) to extend the context window. This approach is feasible since the additional pre-training requires significantly fewer tokens than the initial phase, enabling effective length extrapolation.

Positional encoding is crucial for length extrapolation. Rotary Positional Encoding (RoPE) [8] has become popular due to its superior performance. Typically, RoPE's base frequency is set to 10,000, and models are pre-trained on sequences of 4k tokens. However, an out-of-distribution (OOD) issue can arise when input lengths exceed the original context window without additional measures. To address this, several RoPE variants have emerged, such as PI [2], ABF, NTK [3], and YaRN [5]. Despite differing in form, these variants share the goal of introducing a scaling mechanism to improve performance on extended contexts.

Despite these efforts, it remains unclear which variant is consistently superior or why RoPE is effective in Transformer models. Therefore, our work aims to address the following research questions:

- In long-context scenarios, how do different RoPE-based positional encoding schemes compare in terms of principles and performance?
- What underlying mechanism makes RoPE effective in LLMs?

Ultimately, our goal is to propose a new positional encoding that outperforms existing methods in long-context scenarios.

2 Preliminary

In this section, we present the formal definitions for Attention [11] and Rotary Positional Encoding (RoPE) [8].

Attention operates over a sequence of C embeddings, represented as $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_C]^\top \in \mathbb{R}^{C\times d}$, where d is the model dimension. Learned weight matrices $\mathbf{W}_v \in \mathbb{R}^{d\times d_k}$, $\mathbf{W}_q \in \mathbb{R}^{d\times d_k}$, and $\mathbf{W}_k \in \mathbb{R}^{d\times d_k}$ are applied to these inputs, where d_k is the dimension of the projected embeddings. The attention mechanism computes the attention matrix and uses it to produce a weighted sum of the value vectors, as follows:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}}\right)\mathbf{V}.$$

Submitted to AML course at Tsinghua University, 2024 Fall.

In basic attention, the query, key, and value matrices are computed as $\mathbf{Q} = \mathbf{X}\mathbf{W}_q$, $\mathbf{K} = \mathbf{X}\mathbf{W}_k$, and $\mathbf{V} = \mathbf{X}\mathbf{W}_v$. However, this formulation does not explicitly account for the relative positions of keys and values.

RoPE [8] addresses this by encoding positional information through a phase rotation applied to each element of the embedding vectors. Formally, we define a transformation $\mathbf{f}$ as:

$$\mathbf{f}_{\mathbf{W}}(\mathbf{x}_i, \boldsymbol{\theta}) = \mathbf{R}(\boldsymbol{\theta}, i)\mathbf{W}^\top \mathbf{x}_i,$$

where $\mathbf{x}_i \in \mathbb{R}^{d_k}$ is the embedding at position i, $\mathbf{W}$ is a projection matrix, and $\boldsymbol{\theta} \in \mathbb{R}^{d_k/2}$ is a frequency basis. The rotary transformation matrix $\mathbf{R}(\boldsymbol{\theta}, i)$ is defined as:

$$\mathbf{R}(\boldsymbol{\theta}, i) = \begin{pmatrix} \cos i\theta_1 & -\sin i\theta_1 & \cdots & 0 & 0 \\ \sin i\theta_1 & \cos i\theta_1 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & \cos i\theta_{\frac{d_k}{2}} & -\sin i\theta_{\frac{d_k}{2}} \\ 0 & 0 & \cdots & \sin i\theta_{\frac{d_k}{2}} & \cos i\theta_{\frac{d_k}{2}} \end{pmatrix}.$$

This matrix has the property that $\mathbf{R}(\boldsymbol{\theta}, n-m) = \mathbf{R}(\boldsymbol{\theta}, m)^\top \mathbf{R}(\boldsymbol{\theta}, n)$, based on Ptolemy's identity. As a result, the query-key product between two positions m and n is redefined as:

$$\mathbf{q}_m^\top \mathbf{k}_n = \mathbf{f}_{\mathbf{W}_q}(\mathbf{x}_m, \boldsymbol{\theta})^\top \mathbf{f}_{\mathbf{W}_k}(\mathbf{x}_n, \boldsymbol{\theta}),$$

where the relative positional information $n - m$ is implicitly encoded in the attention score through the query-key interaction.

In the standard RoPE transformation, the components of $\boldsymbol{\theta}$ are defined as $\theta_j = b^{-\frac{2j}{d_k}}$ with a base frequency of $b = 10,000$.

3 Related Work

Positional Encoding. Positional encoding is a crucial component in the transformer architecture [12]. RoPE, introduced by [9], applies sinusoidal rotations to hidden representations before the self-attention mechanism. Numerous RoPE-based improvements have since been developed. ABF [13, 6] involves increasing the base frequency θ in RoPE (e.g., from 10,000 to 500,000), as seen in recent LLaMA 3 models [10]. This adjustment reduces the attention decay effect, enabling the model to handle longer contexts. Other works have highlighted that RoPE's performance degrades in out-of-distribution (OOD) settings. For example, NoPE [4] introduces a causal mechanism to learn positional encodings. Additionally, [7] proposed randomized positional encodings, claiming that they improve OOD performance and enable the model to capture relationships over longer text spans. Recent work [1] provides a comprehensive analysis of existing positional encoding methods and offers explanations for their performance.

4 Initial Methodology

We intend to undertake the following steps:

- **Data recipe and upsampling**: We will systematically design the data recipe and determine an optimal mix of short and long data sequences. This approach ensures that the continuous pre-training phase utilizes a scientifically balanced mixture of data from different domains, facilitating effective long-context extrapolation experiments.
- **Scaling experiments**: We will conduct scaling experiments on models with varying parameter sizes to assess the performance of different positional encoding methods. These experiments will provide a solid foundation for the development of our own positional encoding approach.

Through these steps, we aim to identify the key factors influencing the efficacy of different positional encoding schemes and establish a robust experimental basis for proposing a more effective positional encoding method.

References

[1] Federico Barbero, Alex Vitvitskyi, Christos Perivolaropoulos, Razvan Pascanu, and Petar Veličković. Round and round we go! what makes rotary positional encodings useful?, 2024.

[2] Shouyuan Chen, Sherman Wong, Liangjian Chen, and Yuandong Tian. Extending context window of large language models via positional interpolation. *arXiv preprint arXiv:2306.15595*, 2023.

[3] emozilla. Dynamically Scaled RoPE further increases performance of long context LLaMA with zero fine-tuning, 2023.

[4] Amirhossein Kazemnejad, Inkit Padhi, Karthikeyan Natesan Ramamurthy, Payel Das, and Siva Reddy. The impact of positional encoding on length generalization in transformers, 2023.

[5] Bowen Peng, Jeffrey Quesnelle, Honglu Fan, and Enrico Shippole. Yarn: Efficient context window extension of large language models. *arXiv preprint arXiv:2309.00071*, 2023.

[6] Baptiste Rozière, Jonas Gehring, Fabian Gloeckle, Sten Sootla, Itai Gat, Xiaoqing Ellen Tan, Yossi Adi, Jingyu Liu, Romain Sauvestre, Tal Remez, Jérémy Rapin, Artyom Kozhevnikov, Ivan Evtimov, Joanna Bitton, Manish Bhatt, Cristian Canton Ferrer, Aaron Grattafiori, Wenhan Xiong, Alexandre Défossez, Jade Copet, Faisal Azhar, Hugo Touvron, Louis Martin, Nicolas Usunier, Thomas Scialom, and Gabriel Synnaeve. Code llama: Open foundation models for code, 2024.

[7] Anian Ruoss, Grégoire Delétang, Tim Genewein, Jordi Grau-Moya, Róbert Csordás, Mehdi Bennani, Shane Legg, and Joel Veness. Randomized positional encodings boost length generalization of transformers. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1889–1903, Toronto, Canada, July 2023. Association for Computational Linguistics.

[8] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

[9] Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding, 2023.

[10] Llama 3 Team. The llama 3 herd of models, 2024.

[11] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.

[12] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023.

[13] Wenhan Xiong, Jingyu Liu, Igor Molybog, Hejia Zhang, Prajjwal Bhargava, Rui Hou, Louis Martin, Rashi Rungta, Karthik Abinav Sankararaman, Barlas Oguz, Madian Khabsa, Han Fang, Yashar Mehdad, Sharan Narang, Kshitiz Malik, Angela Fan, Shruti Bhosale, Sergey Edunov, Mike Lewis, Sinong Wang, and Hao Ma. Effective long-context scaling of foundation models. In Kevin Duh, Helena Gomez, and Steven Bethard, editors, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4643–4663, Mexico City, Mexico, June 2024. Association for Computational Linguistics.