Context-aware Rotary Position Embedding

Ali Veisi **Delaram Fartoot** **Hamidreza Amirzadeh**
Axiom Lab
{ali.veisi, h.amirzadeh, d.fartoot}@axiomlab.org

Abstract

Positional encoding is a vital component of Transformer architectures, enabling models to incorporate sequence order into self-attention mechanisms. Rotary Positional Embeddings (RoPE) have become a widely adopted solution due to their compatibility with relative position encoding and computational efficiency. However, RoPE relies on static, input-independent sinusoidal frequency patterns, limiting its ability to model context-sensitive relationships. In this work, we propose CARoPE (Context-Aware Rotary Positional Embedding), a novel generalization of RoPE that dynamically generates head-specific frequency patterns conditioned on token embeddings. This design introduces token- and context-sensitive positional representations while preserving RoPE's efficiency and architectural simplicity. CARoPE computes input-dependent phase shifts using a bounded transformation of token embeddings and integrates them into the rotary mechanism across attention heads. We evaluate CARoPE on the FineWeb-Edu-10B dataset using GPT-2 variants trained on next-token prediction tasks. Experimental results show that CARoPE consistently outperforms RoPE and other common positional encoding baselines, achieving significantly lower perplexity, even at longer context lengths. Additionally, CARoPE enables faster training throughput without sacrificing model stability. These findings demonstrate that CARoPE offers a scalable, expressive, and efficient upgrade to existing positional encoding strategies in Transformer models.

1 Introduction

Transformer architectures have revolutionized the field of deep learning (Vaswani, 2017), achieving state-of-the-art performance across a wide range of tasks in natural language processing (Devlin et al., 2019; Liu, 2019; Chowdhery et al., 2023; Team et al., 2023; Touvron et al., 2023; Achiam et al., 2023). A key component of their success is the self-attention mechanism, which enables the model to dynamically capture relationships between elements in a sequence, regardless of their distance. However, unlike traditional sequence models such as Recurrent Neural Networks (RNNs) (Sherstinsky, 2020) or Convolutional Neural Networks (CNNs) (Gehring et al., 2017), transformers lack an inherent sense of order or position (Yun et al., 2019). This makes *positional encoding* a crucial component, as it injects position-related information into the model to enable sequence-aware processing.

Over the years, several strategies for positional encoding have been proposed. These include fixed sinusoidal embeddings (Vaswani, 2017), learnable absolute position embeddings (Devlin et al., 2019), relative position encodings (Press et al., 2021; Raffel et al., 2020), and rotary positional embeddings (RoPE) (Su et al., 2024). Among these, RoPE has become one of the most widely adopted approaches due to its compatibility with self-attention and ability to encode relative positions through rotation-based transformations.

RoPE works by rotating the query and key vectors within the multi-head attention mechanism using fixed sinusoidal frequencies. Although effective, RoPE still relies on predefined static frequency patterns that are uniform across different inputs and attention heads. As a result, it remains position-dependent but not token- or context-dependent, limiting its expressiveness in modeling more nuanced sequence structures.

In this work, we propose **CARoPE** (*Context-Aware Rotary Positional Embedding*), a novel enhancement of RoPE that introduces *dynamic, input-dependent frequency values* for each attention head. By making frequency generation sensitive to the input content, CARoPE enables the model to adaptively encode positional information in a way that reflects both the position and the underlying context. This results in more expressive and flexible positional representations that are conditioned on the input context and vary across attention heads.

Unlike RoPE's fixed sinusoidal formulation, CARoPE learns a nonlinear transformation of the input embeddings to generate head-specific frequency patterns, which are then integrated into the rotary positional mechanism. This context-aware extension enables richer, token-sensitive position encoding without sacrificing the efficiency and compatibility of the original RoPE framework. We assess the effectiveness of our approach across multiple benchmark datasets, employing GPT-2 variants for the standard next-token prediction task. CARoPE consistently outperforms existing positional encoding methods, including RoPE, and achieves lower perplexity in generated sequences.

2 Proposed Method

We formulate CARoPE as a generalization of Rotary Positional Embedding (RoPE), designed to introduce context-dependent positional modulation within the attention mechanism. While RoPE encodes relative position through fixed sinusoidal rotations, CARoPE replaces these static frequencies with dynamic, token- and head-specific alternatives.

To motivate our method, we first reinterpret standard RoPE through the lens of phase accumulation. In RoPE, the position-dependent rotation applied to each embedding pair is defined as:

$$\phi_i(m) = m \cdot \theta_i,$$

where m is the sequence position and $\theta_i = 10000^{-2i/d}$ is the fixed frequency assigned to the i-th embedding pair in a d-dimensional space. This can be reformulated as a cumulative sum:

$$\phi_i(m) = \sum_{t=1}^{m} \theta_i.$$

Noting that θ_i follows a geometric progression, the phase term becomes:

$$\phi_i(m) = \sum_{t=1}^{m} \theta_1^i = m \cdot \theta_1^i,$$

revealing that each rotational component increases exponentially with dimension.

CARoPE generalizes this formulation by replacing the fixed base frequency θ_1 with a learned, input-dependent function $f(x_t)$, where $x_t \in \mathbb{R}^d$ is the embedding of the token at position t. The generalized phase term becomes:

$$\phi_i^{(h)}(m) = \sum_{t=1}^{m} f(x_t)_h^i,$$

where h indexes the attention head, and $f(x_t)_h \in (0, 1)$ is a learned, bounded scalar frequency specific to head h and token x_t. This formulation maintains the exponential dimension-wise progression of RoPE but allows the frequency to vary across both tokens and heads, yielding context-aware phase accumulation.

The frequency modulation function f is implemented as:

$$f(x_t) = \frac{1}{\text{softplus}(x_t W) + 1},$$

where $W \in \mathbb{R}^{d \times h}$ projects the token embedding to h scalar values, one per head. The softplus activation ensures positivity, while the inverse squashing maps outputs to the interval $(0, 1)$, promoting stability when raised to higher powers.

After computing $\phi_i^{(h)}(m)$ for each position, head, and dimension, we construct sinusoidal components:

$$\cos\left(\phi_i^{(h)}(m)\right), \quad \sin\left(\phi_i^{(h)}(m)\right),$$

which are then applied to the query and key vectors using the standard RoPE formulation.

To preserve stability and enable efficient training, we initialize CARoPE using the standard RoPE formulation. Since RoPE corresponds to a special case of CARoPE. This initialization ensures the model begins with a valid and expressive positional prior.

3 Experiment Setup

3.1 Datasets

For training, we use the FineWeb dataset (Penedo et al., 2024), a large-scale dataset (15 trillion tokens) for LLM pretraining, derived from 96 CommonCrawl snapshots. FineWeb has been shown to produce better-performing LLMs than other open pretraining datasets (Penedo et al., 2024). More specifically, we use a 10B sample of the FineWeb-Edu dataset, which consists of 1.3T tokens from educational web pages filtered from the FineWeb dataset. We allocate 9.9B tokens for training and 0.1B for evaluation. For evaluation, we use the test set of FineWeb-Edu.

GPT-Small models				
Sequence Length	RoPE	CARoPE	Learnable	Sinusoidal
512	21.31	**21.23**	21.90	22.14
1024	56.61	**21.39**	-	166.18
GPT-Tiny models				
Sequence Length	RoPE	CARoPE	Learnable	Sinusoidal
512	29.33	**28.99**	30.48	30.62
1024	81.27	**36.74**	-	223.28

Table 1: Perplexity comparison on the FineWeb-Edu-10B evaluation set. The first row reports results from GPT-Small models, and the second row shows results from GPT-Tiny models. All models were trained for 19k steps on the FineWeb-Edu-10B training set with a context length of 512.

3.2 Settings

For all next-token prediction tasks, we use the GPT-2 variants (Brown et al., 2020). For the FineWeb-Edu-10B dataset, we use its small version (12 layers, 10 heads, and a hidden dimension of 768) with 124M parameters, and a tiny version of GPT-2 (44M parameters) with 6 layers, 8 heads, and a hidden dimension of 512. The evaluation metric is perplexity (PPL), and we train the models with sequence length of 512. All the models are trained on two H100 GPUs with 80G GPU RAM. Training settings are the same as those used for GPT-2 (Radford et al., 2019). Gradients are updated after processing 524,288 tokens and vocab size is 50304. For training on the FineWeb-Edu-10B dataset, we run 19k steps (~1 epoch) with batch sizes of 64, and 32 for the tiny, and small models, respectively. The learning rate starts at 0.0006, with a linear warmup over 750 steps, followed by cosine decay to a minimum of 0.00006.

3.3 Baselines

We compare our method against the following positional encoding approaches:

Learnable (Vaswani, 2017): A trainable additive positional encoding (APE) where each position is associated with a learned embedding. The number of positions is fixed and predefined during training.

Sinusoidal (Vaswani, 2017): A fixed APE used in early Transformer models (Vaswani, 2017; Baevski and Auli, 2018; Ott et al., 2018; Lewis et al., 2021).

RoPE (Su et al., 2024): A non-learnable relative positional encoding (RPE) widely adopted in LLMs such as GPT-2 (Brown et al., 2020), LLaMA (Touvron et al., 2023), PaLM (Chowdhery et al., 2023), and Gemma (Team et al., 2024a,b).

4 Results

Table 1 reports the perplexity of models trained with sequence length of 512 and different positional encoding strategies on the FineWeb-Edu-10B evaluation set. Across both GPT-Small and GPT-Tiny variants, CARoPE consistently outperforms RoPE, achieving notably lower perplexity, especially at longer sequence lengths. For example, at a sequence length of 1024, CARoPE reduces perplexity by more than 60% compared to RoPE in the GPT-Tiny model (36.74 vs. 81.27). This demonstrates CARoPE's ability to generalize better over longer contexts.

The results validate the effectiveness of dynamic, input-dependent frequency modulation in enhancing positional representation. Notably, CARoPE not only achieves better perplexity but also enables faster training, processing approximately 0.76 million tokens per second compared to 0.63 million for RoPE in GPT-Small models.

5 Conclusion

We presented CARoPE, a context-aware extension of Rotary Positional Embeddings that introduces input- and head-dependent frequency modulation. By dynamically adapting to token content, CARoPE improves the expressiveness of positional encoding with minimal overhead. Our experiments demonstrate consistent gains over RoPE across model sizes and sequence lengths, highlighting its effectiveness for enhancing Transformer-based language models.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Alexei Baevski and Michael Auli. 2018. Adaptive input representations for neural language modeling. *arXiv preprint arXiv:1809.10853*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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)*, pages 4171–4186.

Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N Dauphin. 2017. Convolutional sequence to sequence learning. In *International conference on machine learning*, pages 1243–1252. PMLR.

Mike Lewis, Shruti Bhosale, Tim Dettmers, Naman Goyal, and Luke Zettlemoyer. 2021. Base layers: Simplifying training of large, sparse models. In *International Conference on Machine Learning*, pages 6265–6274. PMLR.

Yinhan Liu. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 364.

Myle Ott, Sergey Edunov, David Grangier, and Michael Auli. 2018. Scaling neural machine translation. *arXiv preprint arXiv:1806.00187*.

Guilherme Penedo, Hynek Kydlíček, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, Thomas Wolf, et al. 2024. The fineweb datasets: Decanting the web for the finest text data at scale. *arXiv preprint arXiv:2406.17557*.

Ofir Press, Noah A Smith, and Mike Lewis. 2021. Train short, test long: Attention with linear biases enables input length extrapolation. *arXiv preprint arXiv:2108.12409*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

Alex Sherstinsky. 2020. Fundamentals of recurrent neural network (rnn) and long short-term memory (lstm) network. *Physica D: Nonlinear Phenomena*, 404:132306.

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

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. 2024a. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. 2024b. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

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

Chulhee Yun, Srinadh Bhojanapalli, Ankit Singh Rawat, Sashank J Reddi, and Sanjiv Kumar. 2019. Are transformers universal approximators of sequence-to-sequence functions? *arXiv preprint arXiv:1912.10077*.