3D-RPE: Enhancing Long-Context Modeling Through 3D Rotary Position Encoding

Xindian Ma[1], Wenyuan Liu[1], Peng Zhang[1*], and Xu Nan[2]

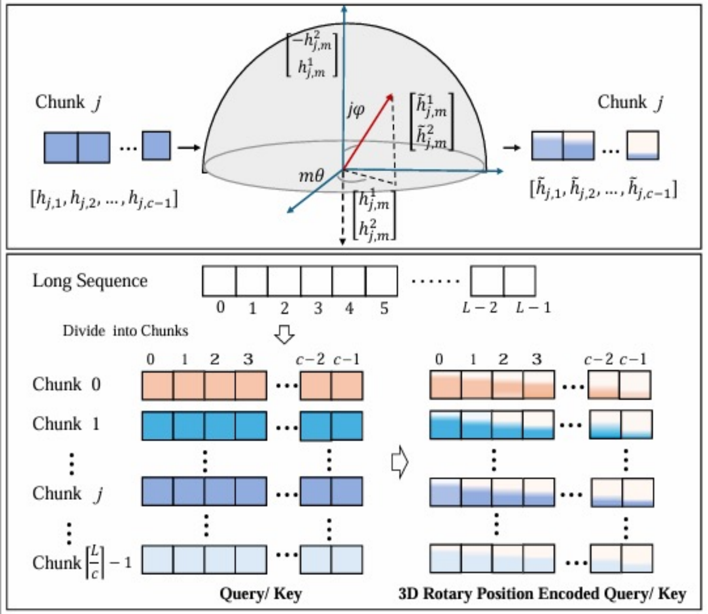
AAAI2025

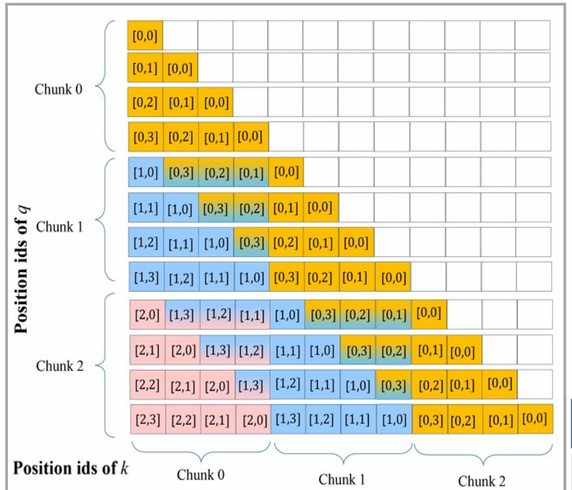
The 39th Annual AAAI Conference on Artificial Intelligence

Contribution

- A position encoding method on a 3D sphere, 3D-RPE, is provided, which can enhance the long-context modeling capability of LLMs by replacing RoPE.
- It is proved that 3D-RPE has two benefits, controllable long-term decay and mitigating the reduction in positional resolution.
- LLMs combine with 3D-RPE have achieved significant performance improvements in long-context NLU tasks.

Methodology





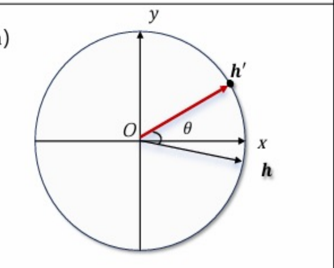
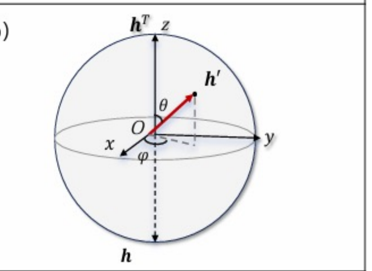
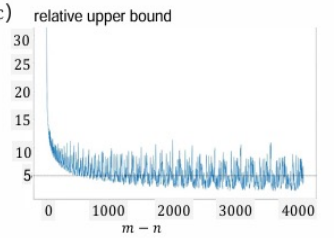
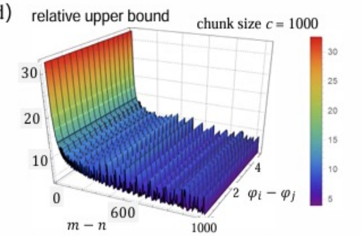
Visualization of **the relative position matrix A** employing 3D-RPE, with chunk size 4, and sequence size L=12.

Definition (3D Rotary Position Encoding). Let $\boldsymbol{h}_{j,m} \in \mathbb{R}^d$ be a state vector of an attention head without position encoding, where d is the dimension of the vector, which is an even number. **3D-RPE encodes $\boldsymbol{h}_{j,m}$ into the vector $\widetilde{\boldsymbol{h}}_{j,m}$**, which is formalized as:

$$\widetilde{\boldsymbol{h}}_{j,m} = e^{-im\theta}(\cos\varphi_j \boldsymbol{h}_{j,m}^{\perp} + \sin\varphi_j \boldsymbol{h}_{j,m})$$

where i is the imaginary unit, and $\boldsymbol{h}_{j,m}^{\perp}$ equals to $[-\boldsymbol{h}_{j,m}^2, \boldsymbol{h}_{j,m}^1]^T$.

Benefits

Method	2D Rotary Position Encoding (RoPE)	3D Rotary Position Encoding(3D-RPE)
Schematic Drawing	(a) 	(b) 
Formula	$f_{\{q,k\}}(\boldsymbol{h}, m) = e^{im\theta}\boldsymbol{h}$	$f_{\{q,k\}}(\boldsymbol{h}, m, j) = e^{im\theta}(\cos\varphi_j \boldsymbol{h}^{\perp} + \sin\varphi_j \boldsymbol{h})$
Long-term Decay	(c) 	(d) 
Position Resolution	(e) PI $\varepsilon_{rope} = 1 \longrightarrow \varepsilon'_{rope} = \frac{L_p}{L}$	(f) PI $\varepsilon_{3d-rpe} = 1 \longrightarrow \varepsilon'_{3d-rpe} > \frac{L_p}{L}$

$$|s(\boldsymbol{q}_{i,m}, \boldsymbol{k}_{j,n}, \boxed{\varphi_i - \varphi_j}, m-n)| \le |e^{i(\varphi_i-\varphi_j)}|.$$

$$\left|\sum_{l=0}^{\frac{d}{2}-1} E_{l+1}(h_{l+1}-h_l)\right| \le (\max_l |h_{l+1}-h_l|) \sum_{l=0}^{d/2-1} |E_{l+1}|$$

By introducing positional modeling on chunks, the mitigation of long-term decay is achieved.

Theorem (Improved Position Resolution). *For a pre-trained language model with a length of L_p and an extension length requirement of L, employing linear position interpolation extension methods $\mathcal{I}$ based on Rotary Position Encoding (RoPE) can elevate the relative position resolution from $\mathcal{E}_{rope}$ to $\mathcal{E}'_{rope}$. Let $\mathcal{E}'_{3d-rpe}$ denote the relative positional encoding resolution achieved by the method $\mathcal{I}$ based on 3D-RPE, with chunk size $c \ge 3$, there is:*

$$\mathcal{E}'_{3d-rpe} > \mathcal{E}'_{rope}$$

Theoretically, it is proven that when the chunk size is greater than 3, the **positional interpolation resolution of 3D-RPE is greater than that of RoPE.**

Experimental Results

METHODS	Single-Doc QA	Multi-Doc QA	Summarization	Few-shot	Code
LLaMA-2-7B-chat	24.90	22.60	24.70	60.01	48.10
LLaMA-2-7B-chat-PI	18.98	17.16	25.03	49.43	52.73
LLaMA-2-7B-chat-NTK	23.21	23.34	24.40	59.29	49.28
StreamingLLM	21.47	22.22	22.20	50.05	48.00
ChunkLLaMA-16k	24.04	22.98	21.52	46.31	49.73
LongChat-32k	31.58	23.50	26.70	64.02	54.10
LongAlpaca-16k	28.70	28.10	27.80	63.70	56.00
LongLLaMA	30.12	16.37	24.19	60.31	66.05
Vicuna-v1.5-7B-16k	28.01	18.63	26.01	66.20	47.30
ChatGLM3-6B-32k	40.30	46.60	**29.50**	68.10	56.20
3D-RPE-LLaMA2-7B-Chat	**47.40**	**60.10**	28.99	**73.16**	**76.50**

Code: https://github.com/maxindian/3D-RPE Long-Contex-Modeling

Conclusion

We present a novel rotary position encoding method called 3D-RPE. Compared to RoPE, we have theoretically proved that 3D-RPE possesses two key advantages: controllable long-term decay and improved interpolation resolution. Experimentally, 3D-RPE has excelled in long-context NLU tasks.

Contact: xindianma@tju.edu.cn