Wavelet-based Positional Representation for Long Context

Yui Oka Taku Hasegawa Kyosuke Nishida Kuniko Saito

NTT Human Informatics Laboratories, NTT Corporation

Overview of the Task: Long Context and Extrapolation

- ☐ Position encoding focuses on representation using sine waves
- ☐ RoPE using sine waves does not have extrapolation performance
- ☐ Extrapolation-capable ALiBi limits the receptive field of attention
- ☐ We propose a new position encoding based on wavelets that is extrapolation-capable without limiting the receptive field of attention

Preliminary

- **Wavelet** is a wave that decays quickly and locally as it approaches zero. The wavelet function ψ is defined as follows.In this case, b is the shift and $a > 0$ is the scale parameter.

$$\psi_{a,b}(t) = \frac{1}{\sqrt{a}}\psi\Big(\frac{t-b}{a}\Big).$$

- **Wavelet transform (WT)** is the process of transforming a signal $x(t)$ into the frequency domain and time domain by computing the inner product of the wavelet function $\psi_{a,b}(t)$ and signal $x(t)$. When $a \in [2,4]$ and $b \in [0,1,2,3]$, the wavelet transform can be expressed in terms of determinants as follows:

$$\begin{bmatrix} W(2,0) \\ W(4,0) \\ W(2,1) \\ W(4,1) \\ \vdots \\ W(4,3) \end{bmatrix} = \begin{bmatrix} \psi_{2,0}(0) & \psi_{2,0}(1) & \psi_{2,0}(2) & \dots & \psi_{2,0}(T-1) \\ \psi_{4,0}(0) & \psi_{4,0}(1) & \psi_{4,0}(2) & \dots & \psi_{4,0}(T-1) \\ \psi_{2,0}(-1) & \psi_{2,0}(0) & \psi_{2,0}(1) & \dots & \psi_{2,0}(T-2) \\ \psi_{4,0}(-1) & \psi_{4,0}(0) & \psi_{4,0}(1) & \dots & \psi_{4,0}(T-2) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \psi_{4,0}(-3) & \psi_{4,0}(-2) & \psi_{4,0}(-1) & \dots & \psi_{4,0}(T-3) \end{bmatrix} \begin{bmatrix} x(0) \\ x(1) \\ x(2) \\ \vdots \\ x(T-1) \end{bmatrix}.$$

- **RoPE** incorporates positional information directly into the self-attention mechanism by rotating the query and key vectors in complex space. When divided into even and odd dimensions, the following calculations are performed for the m-th query in each sequence. In even dimensions, RoPE is expressed as follows.

$$\begin{bmatrix} q_0^m \\ q_2^m \\ \vdots \\ q_{d-2}^m \end{bmatrix} = \begin{bmatrix} \cos m\theta_1 & -\sin m\theta_1 & 0 & 0 & \dots & 0 & 0 \\ 0 & 0 & \cos m\theta_2 & -\sin m\theta_2 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \dots & \cos m\theta_{d/2} & -\sin m\theta_{d/2} \end{bmatrix} \begin{bmatrix} q_0^m \\ q_1^m \\ \vdots \\ q_{d-2}^m \\ q_{d-1}^m \end{bmatrix}.$$

where $q^m \in \mathbb{R}^{1\times d}$ is the m-th query when the number of dimensions is d and $\theta_i = 10000^{-2(i-1)/d}, i \in [1,2,...,d/2]$.

Findings 1: Multi-Window Characteristics in ALiBi

The attention map shows that **ALiBi uses multiple window sizes corresponding to relative positions** and that the window size increases as the slope decreases.

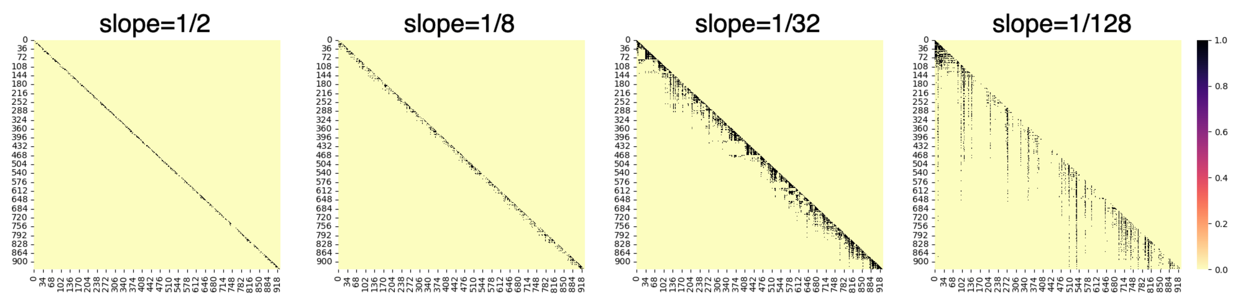


Figure 1. Heatmap of scaled attention scores via softmax normalization in ALiBi without non-overlapping inference. The trainning length is 512, and the inference length is 1012.

Findings 2: RoPE is Wavelet-Transform

First, we show the wavelet transform using the following two Haar-like wavelets.

$$\psi(t) = \begin{cases} \cos f(t) & 0 \le t < 1, \\ -\sin f(t) & 1 \le t < 2, \\ 0 & \text{otherwise.} \end{cases} \quad \psi'(t) = \begin{cases} \sin f(t) & 0 \le t < 1, \\ \cos f(t) & 1 \le t < 2, \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

Assuming that when $x(t)(0 \le t \le d-1)$ is a signal with d elements, the wavelet ψ is used and wavelet transform is performed at each scale $a = 1$. We define the shift parameter as $b_j = j - \delta(j)(j = 0,2,..,d-2)$. Here, $\delta(t)$ is a function such that $0 \le t \le d-1$ and $0 \le \delta(t) < 1$. When the wavelet function is Haar-like wavelet $\psi(t)$ in Eq.(1) and $a = 1$ and $b \in [b_0, b_2, .., d_{d-2}]$, the wavelet matrix ψ in the wavelet transform $w = \psi x$ can be expressed in terms of determinants as follows.

$$\begin{bmatrix} W(1,b_0) \\ W(1,b_2) \\ \vdots \\ W(1,b_{d-2}) \end{bmatrix} = \begin{bmatrix} \cos\phi_0 & -\sin\phi_1 & 0 & 0 & \dots & 0 & 0 \\ 0 & 0 & \cos\phi_2 & -\sin\phi_3 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \dots & \cos\phi_{d-2} & -\sin\phi_{d-1} \end{bmatrix} \begin{bmatrix} x(0) \\ x(1) \\ \vdots \\ x(d-2) \\ x(d-1) \end{bmatrix}. \tag{2}$$

To simplify the notation in the matrix representation above, we write ϕ_j for $j = 0,1,\dots,d-1$, where $\phi_j = f(1+\delta(j))$ if j is odd, and $\phi_j = f(\delta(j))$ otherwise. Let x be the query q^m, and define f such that $\phi_j = \phi_{j+1} = m\theta_{\lceil \frac{j+1}{2} \rceil}$ for $j = 0,2,4,\dots,d-2$, where $\theta_i = 10000^{-2(i-1)/d}$ and $i \in [1,2,...,d/2]$.

$$\begin{bmatrix} W(1,b_0) \\ W(1,b_2) \\ \vdots \\ W(1,b_{d-2}) \end{bmatrix} = \begin{bmatrix} \cos m\theta_1 & -\sin m\theta_1 & 0 & 0 & \dots & 0 & 0 \\ 0 & 0 & \cos m\theta_2 & -\sin m\theta_2 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \dots & \cos m\theta_{d/2} & -\sin m\theta_{d/2} \end{bmatrix} \begin{bmatrix} x(0) \\ x(1) \\ \vdots \\ x(d-2) \\ x(d-1) \end{bmatrix}. \tag{3}$$

RoPE can be viewed as a **wavelet transform using Haar-like wavelets that change amplitude on a fixed scale.** This wavelet transform in RoPE is performed across the number of query head dimensions d.

Motivation

- **Position-based Transformation**: RoPE predominantly relies on independent transformation based on the 'head' dimensions. We apply a wavelet transform based on **the relative position of the sentence**.
- **Type of Wavelet**: RoPE can be thought of as a wavelet transform using the Haar wavelet. We use **more complex wavelet shapes**.
- **Diversification of Window Sizes (Scale Parameters)**: ALiBi have multiple windows and it may effective for long contexts. We introduce **a variety of scale and shift parameters**.

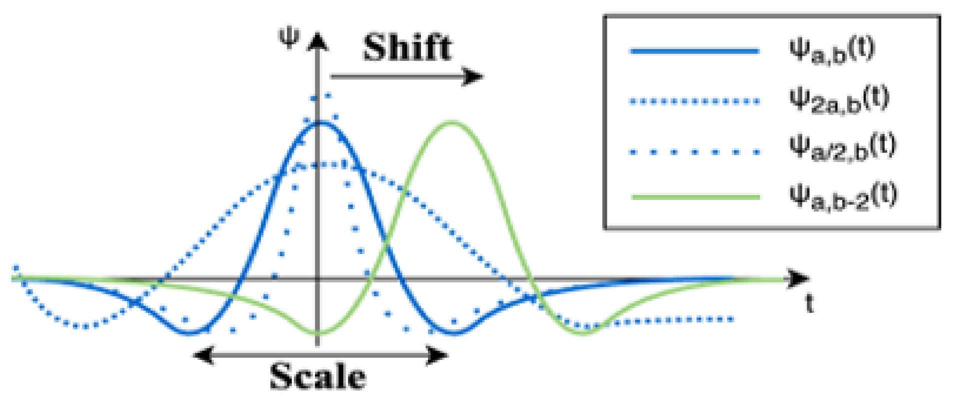


Figure 2. Example of Wavelet

Proposed Method: Wavelet-based Positional Representation

1. Incorporating Wavelet Transform into PE We incorporate wavelets based on RPE.

$$e_{m,n} = \frac{q_m k_n^T + q_m (p_{m,n})^T}{\sqrt{d}}, \tag{4}$$

2. Wavelet Function We used the Ricker wavelet as a base wavelet. We substitute the relative position m-n into t.

$$\psi(t) = (1-t^2)\exp\Big(\frac{-t^2}{2}\Big). \tag{5}$$

3. Shift and scale parameters We use s distinct patterns for the scale parameter a and $\frac{d}{s}$ patterns for the shift parameter b.

$$(a,b) \in \{2^0, 2^1, 2^2, ...2^{s-1}\} \times \{0,1,2,3,...,\frac{d}{s}-1\}. \tag{6}$$

Finally, $p_{m,n}$ is computed as follow

$$p_{m,n} = \Big(1 - \Big(\frac{m-n-b}{a}\Big)^2\Big)\exp\Big(-\frac{1}{2}\Big(\frac{m-n-b}{a}\Big)^2\Big). \tag{7}$$

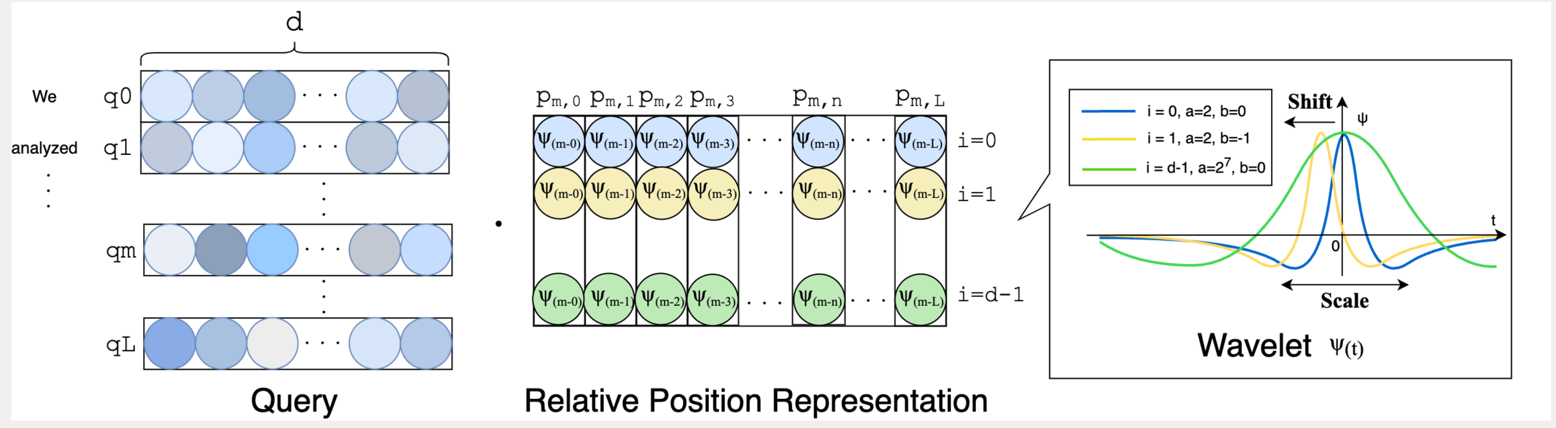


Figure 3. **Overview of Wavelet-based Relative Positional Representation As in RPE (shaw+, 2018)**, our method computes a relative positional representation $(p_{m,n})^T$ to the query q_m and the key k_n. Instead of learnable embedding in RPE, the position is computed based on the wavelet function. Different wavelet functions $\psi_{a,b}$ are used for each dimension of the head d. Furthermore, the scale parameter a and the shift parameter b change depending on the dimension of the head d.

Experiments

Experimental Settings

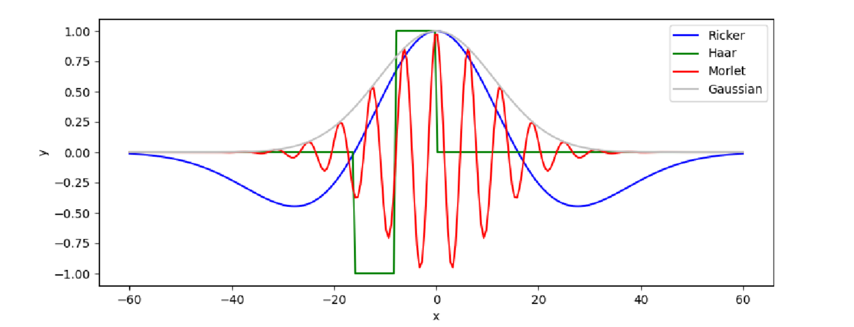
- Model size - d_{emb} is 1024, Head size is 8, d_{head} is 128, layer size is 16.
- Max Allowable Length in Pre-training - 512
- Dataset - Wikitext-103 (Train, Dev, Test)
- Evaluation - Perplexity without Sliding Window

Comparison Method

- ALiBi
- RoPE ($\theta = 10000$ or 500000)
- NoPE - Position information is not given
- Transformer-XL PE - A relative PE that uses sine waves

Wavelet Type in Our method

- Ricker
- Haar
- Morlet
- Gaussian



*We use $s = 8$ scale variants ($a \in \{2^0, 2^1, ..., 2^7\}$) and 16 shift variants ($b \in \{0,1,2,...,15\}$), resulting in $8 \times 16 = 128$ unique wavelets.

Perplexity Results

PE Type	pos	128	256	512	1012	1512	2512
NoPE	-	26.38	23.23	21.53	21.03	21.58	48.48
RoPE ($\theta = 10000$)	abs	23.82	20.98	19.39	23.25	44.38	93.94
RoPE ($\theta = 500000$)	abs	23.81	20.95	19.35	23.70	40.39	77.90
Trans-XL	rel	24.16	21.53	19.96	19.09	18.92	19.05
ALiBi	rel	24.18	21.32	19.69	18.71	18.42	18.41
Wavelet(Ricker)	rel	**23.64**	**20.82**	**19.19**	**18.23**	**18.00**	**17.99**
Haar	rel	23.73	20.89	19.27	18.34	18.11	18.17
Morlet	rel	24.15	21.28	19.65	19.02	20.46	26.56
Gaussian	rel	23.77	20.90	19.30	18.31	18.02	**17.88**

! RoPE cannot be extrapolated without using a sliding window mechanism.

! ALiBi underperforms compared to RoPE on short sentences.

! Our wavelet-based methods—particularly those using the Ricker wavelet—consistently achieve the best performance across all sequence lengths, and they are also naturally extrapolatable.

! The Morlet wavelet, which closely resembles a sine wave in its shape, yielded the poorest results. This suggests that sine-wave-like patterns are not well-suited for encoding relative positional information.

Experiments with Llama-7B

Experimental Settings

- Model size - Llama2-7B
- Max Allowable Length in Pre-training - 4096
- Dataset - Redpajame (Train, Dev) CodeParrot (Test)
- Evaluation - Perplexity with Sliding Window

Perplexity Results

PE Type	pos	4k	8k	16k	32k
RoPE ($\theta = 500000$)	abs	9.45	9.33	9.12	8.90
Wavelet(Ricker)	rel	9.00	9.01	8.83	8.60

*We use $s = 8$ scale variants ($a \in \{2^2, 2^3, ..., 2^9\}$) and 16 shift variants ($b \in \{0,1,2,...,15\}$), resulting in $8 \times 16 = 128$ unique wavelets.

yui.oka@ntt.com