



Wavelet-based Positional Representation for Long Context (ウェーブレット位置符号化)

NLPコロキウム 2025年4月16日

岡 佑依

NTT 人間情報研究所

岡佑依 Yui Oka

■ 学歴

愛媛大（学部） システムトレードの研究

NAIST（修士） NLP、特に位置符号化の研究

東北大（現在） NLP、特に位置符号化の研究

■ 所属：NTT CS研 ⇒ NTT人間研(2024年7月から異動)



長谷川拓 Taku Hasegawa

■ 学歴

大阪府立大（中退） 進化計算

大阪府立大（修士・博士・PD） 最適化アルゴリズムと機械学習

Freiburg 大学（Invited Researcher） NNのアーキテクチャサーチ

■ 所属：NTT人間研

基盤モデル構築・知識転移



背景 - 位置符号化

位置符号化 - Positional Encoding

- Transformerモデル内部で位置を表現する機構の総称

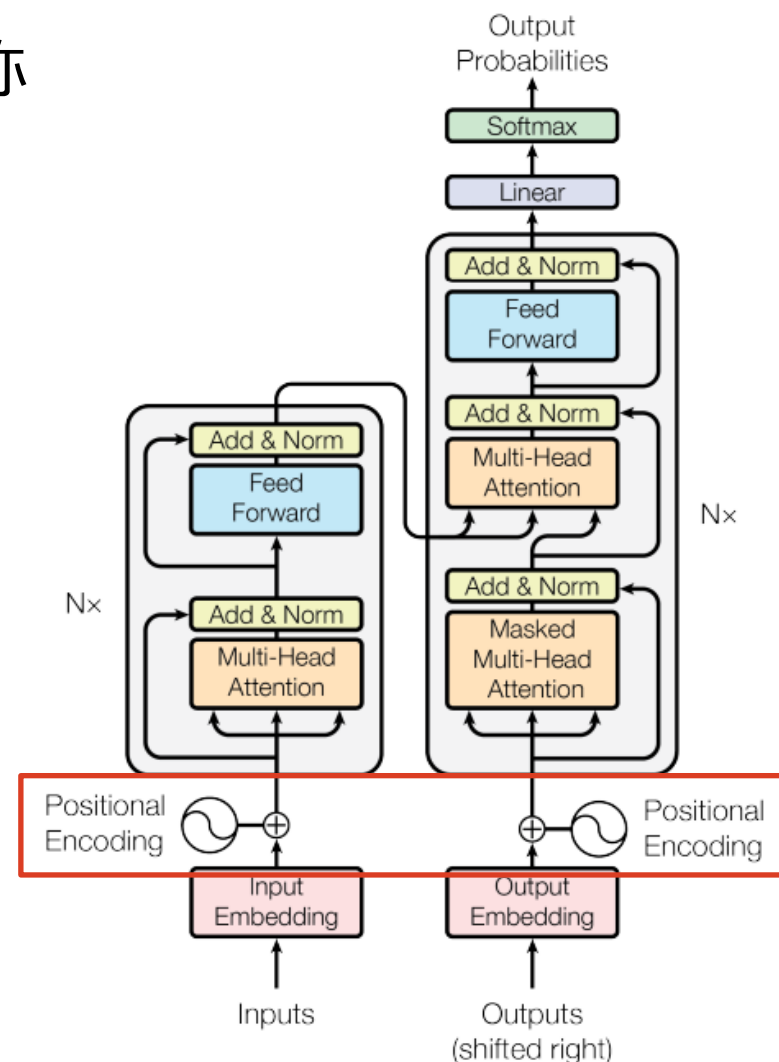
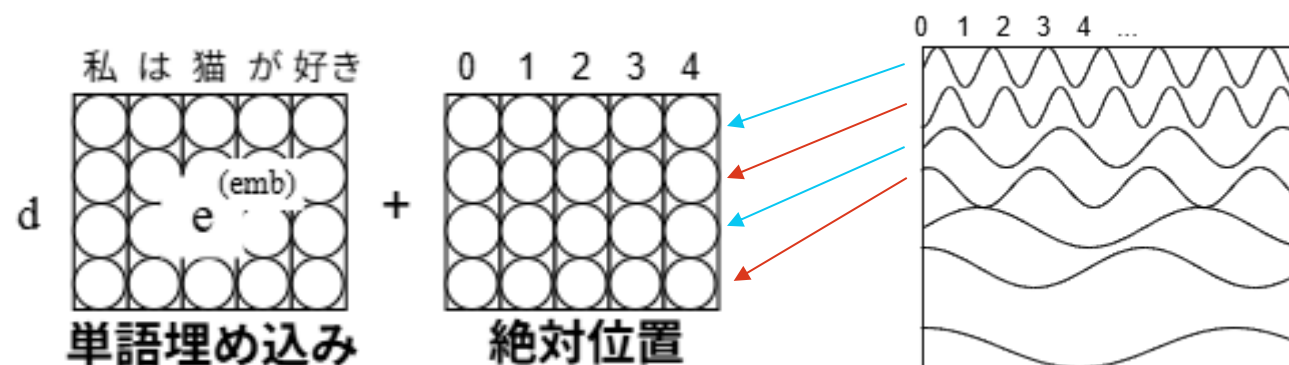
*Position Embeddingとも言われたりします

正弦波位置符号化 Sinusoidal Positional Encoding

- 単語埋め込みにsinまたはcosから計算した値を足す

$$PE_{(pos, 2i)} = \sin(pos/10000^{2i/d_{model}})$$

$$PE_{(pos, 2i+1)} = \cos(pos/10000^{2i/d_{model}})$$

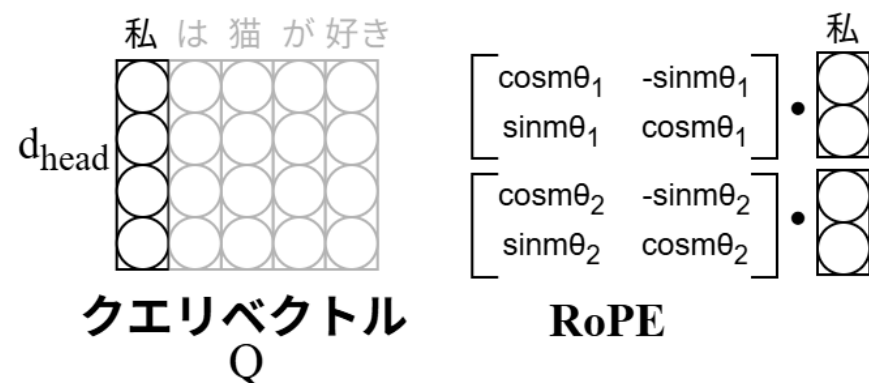


RoPE – Rotary Position Embedding

- クエリ（またはキー）を2次元ごとに分割し、回転行列と内積を取ることで位置を表現する
- 各クエリを回転させていると解釈できる
- θ （回転率）の決め方は正弦波位置符号化に由来

$$\theta_i = 10000^{-2(i-1)/d}, i \in [1, 2, \dots, d/2]$$

$$\begin{bmatrix} \cos\theta_1 & -\sin\theta_1 & 0 & 0 & \dots & 0 & 0 \\ \sin\theta_1 & \cos\theta_1 & 0 & 0 & \dots & 0 & 0 \\ 0 & 0 & \cos\theta_2 & -\sin\theta_2 & \dots & 0 & 0 \\ 0 & 0 & \sin\theta_2 & \cos\theta_2 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \dots & \cos\theta_{d/2} & -\sin\theta_{d/2} \\ 0 & 0 & 0 & 0 & \dots & \sin\theta_{d/2} & \cos\theta_{d/2} \end{bmatrix} \begin{bmatrix} q_0^m \\ q_1^m \\ q_2^m \\ q_3^m \\ \vdots \\ q_{d-2}^m \\ q_{d-1}^m \end{bmatrix}$$



- 多くのLLMで採用されている
 - LlamaやPhiモデルでは、事前学習の段階で θ の拡張が行われている
 - コンテキスト拡張のための再学習でも θ の拡張（Position Interpolation）を行っている

本日は触れません 😞

■ Attention is All You Need[1]から引用

In this work, we use sine and cosine functions of different frequencies:

$$PE_{(pos, 2i)} = \sin(pos/10000^{2i/d_{\text{model}}})$$
$$PE_{(pos, 2i+1)} = \cos(pos/10000^{2i/d_{\text{model}}})$$

where pos is the position and i is the dimension. That is, each dimension of the positional encoding corresponds to a sinusoid. The wavelengths form a geometric progression from 2π to $10000 \cdot 2\pi$. We chose this function because we hypothesized it would allow the model to easily learn to attend by relative positions, since for any fixed offset k , PE_{pos+k} can be represented as a linear function of PE_{pos} .

相対位置も学習できるんじゃないか

We also experimented with using learned positional embeddings [9] instead, and found that the two versions produced nearly identical results (see Table 3 row (E)). We chose the sinusoidal version because it may allow the model to extrapolate to sequence lengths longer than the ones encountered during training.

学習時の最大長よりも長い文を外挿可能なんじゃないか

■ RoFormer: Enhanced Transformer with Rotary Position Embedding^[2]から引用

In this paper, we introduce a novel method, namely Rotary Position Embedding(RoPE), to leverage the positional information into the learning process of PLMS. Specifically, RoPE encodes the absolute position with a rotation matrix and meanwhile incorporates the explicit relative position dependency in self-attention formulation. Note that the proposed RoPE is prioritized over the existing methods through valuable properties, including the sequence length flexibility, decaying inter-token dependency with increasing relative distances, and the capability of equipping the linear self-attention with relative position encoding. Experimental results on various long text classification benchmark datasets show that the enhanced transformer with rotary position embedding, namely RoFormer, can give better performance compared to baseline alternatives and thus demonstrates the efficacy of the proposed RoPE.

回転行列によってSelf-attentionの計算に相対位置を組み込める

正弦波位置符号化とRoPEは相対位置を表現可能か？

■ 相対位置とは？

- 各トークンから各トークンに対する位置
- RPE[2]やALiBi[3]、Transformer-XL[4]などの手法がある

	私	は	猫	が	好き
絶対位置	0	1	2	3	4

	私	は	猫	が	好き
私	0	1			
は	-1	0	1		
猫		-1	0	1	
が			-1	0	1
好き				-1	0
相対位置					

■ 相対位置が取れるという理論的証明

- 正弦波位置符号化

$$\begin{bmatrix} \cos(\frac{k}{\theta}) & \sin(\frac{k}{\theta}) \\ -\sin(\frac{k}{\theta}) & \cos(\frac{k}{\theta}) \end{bmatrix} \begin{bmatrix} \sin(\frac{pos}{\theta}) \\ \cos(\frac{pos}{\theta}) \end{bmatrix} = \begin{bmatrix} \sin(\frac{pos+k}{\theta}) \\ \cos(\frac{pos+k}{\theta}) \end{bmatrix}$$

この回転行列を使ってアフィン変換によって表現していると捉えることができる

- RoPE

$$\mathbf{q}_m^\top \mathbf{k}_n = (\mathbf{R}_{\Theta, m}^d \mathbf{W}_q \mathbf{x}_m)^\top (\mathbf{R}_{\Theta, n}^d \mathbf{W}_k \mathbf{x}_n) = \mathbf{x}^\top \mathbf{W}_q \mathbf{R}_{\Theta, n-m}^d \mathbf{W}_k \mathbf{x}_n$$

RoPEを適用したあと、クエリとキーの内積を計算すると、n-mの相対位置を計算した式と一致する

A. 理論的には表現可能

本日は詳しく触れません

- [2] <https://arxiv.org/abs/1803.02155>
[3] <https://openreview.net/forum?id=R8sQPpGCv0>
[4] <https://arxiv.org/abs/1901.02860>

正弦波位置符号化とRoPEには外挿機能があるのか？

■ 外挿とは？

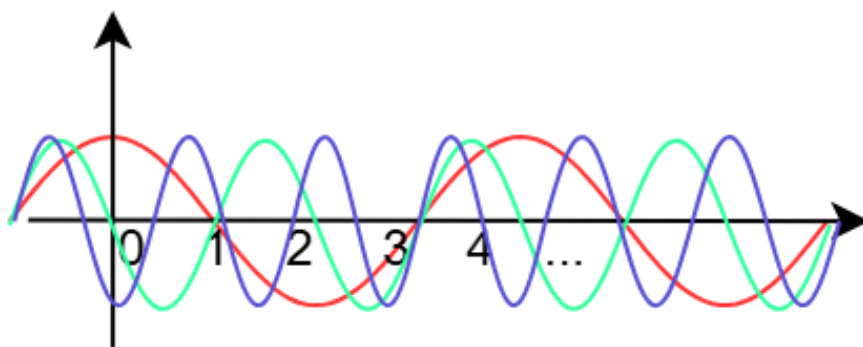
- （機械学習の文脈では）学習したデータの範囲外のデータで予測を行うこと
- **学習時の最大長を超えた文を生成すること**

学習時に決められた最大長

吾輩は猫である。名前はまだ無い。どこで生れたかとうんと見当がつかぬ。

外挿

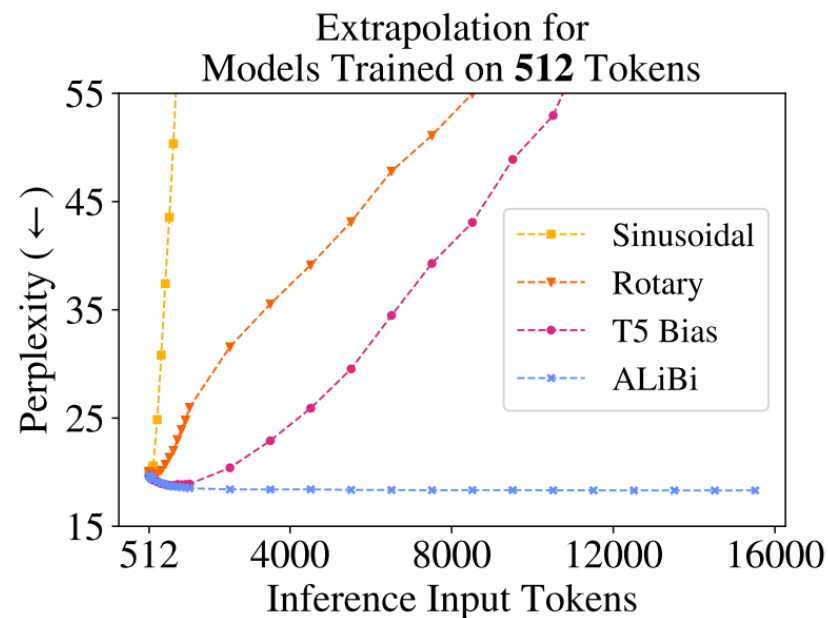
 **正弦波は無限に波が続くので確かに外挿と相性がよさそうに思える**



正弦波位置符号化とRoPEには外挿機能があるのか？

■ Train Short, Test Long: Attention with Linear Biases Enables Input Length Extrapolation[5]

- 正弦波位置符号化などの外挿性能を調査
- Perplexityが発散するので外挿性能は低い

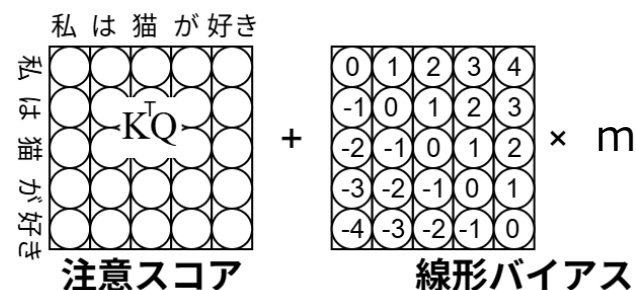


- 外挿性能が最も高いALiBiを提案

● ALiBi (Attention with Linear Biases)

- 注意スコアに線形バイアスを足す
- 傾きmはヘッド毎に異なる ($m=1/2, 1/4, \dots, 1/256$)

$$\text{softmax}(\mathbf{q}_i \mathbf{K}^T + m \cdot [-(i-1), \dots, -2, -1, 0]),$$



A. 正弦波位置符号化とRoPEに外挿機能はおそらくない，相対位置だと外挿できる

(補足) ALiBi使えばよくない？

A. ALiBiはAttentionの受容野を制限するからあんまりよくない

■ Dissecting Transformer Length Extrapolation via The Lens of Receptive Field Analysis [6]

- ALiBiを使うとAttentionの受容野が制限されSliding Window Attentionのようになる
- 遠い依存関係にある単語の情報が得られない可能性がある

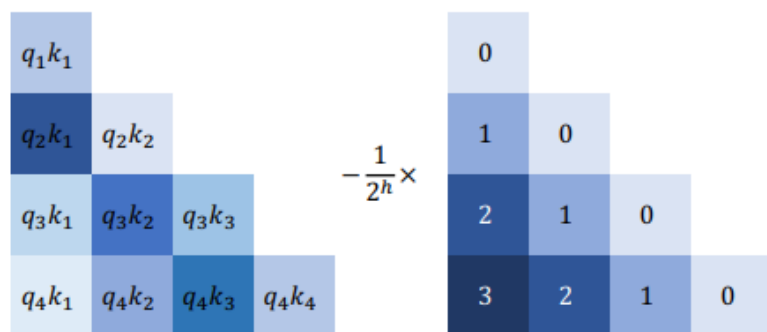


Figure 1: ALiBi. For a transformer language model with H attention heads, the range of h is $n \cdot \frac{8}{H}$, where $n = \{1 \dots H\}$. Left = self-attention matrix, right = temporal biases matrix.



Figure 2: Windowed Attention. This is the same design as Longformer (Beltagy et al., 2020). We limit the context window size to $w = 2$ in this example. Left = self-attention matrix, right = temporal biases matrix.

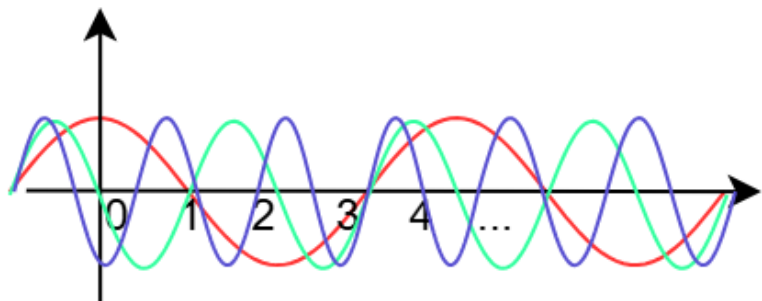
目的・アイデア

目的 - 外挿もできて性能がいい位置符号化が欲しい！



現状

- 外挿可能だが、Attentionの受容野を制限している (ALiBi)
- 正弦波を使った位置符号化が多いが、外挿は不可能(RoPEや正弦波PE)
 - 相対位置を表現できる、外挿ができる、と言われているが・・・
 - 複数の周波数を組み合わせる、という**フーリエ変換**のような処理をしている？



🤔 もし、相対位置が活用できているなら
外挿できるんじゃないか？

🤔 外挿できないのに正弦波を使うべきなのか？

フーリエ変換と自然言語のギャップ

■ フーリエ変換

- 周波数解析の一種
- 信号を周波数領域へと変換
- 定常的な信号の解析に有効

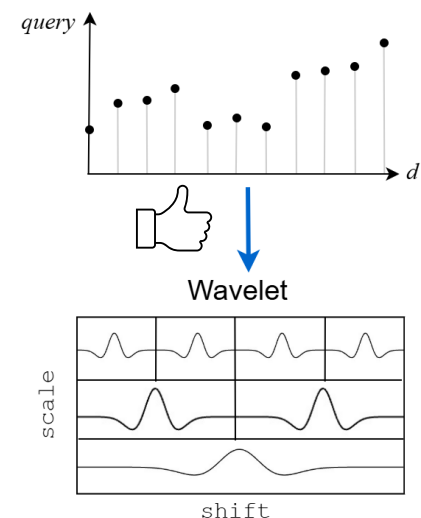
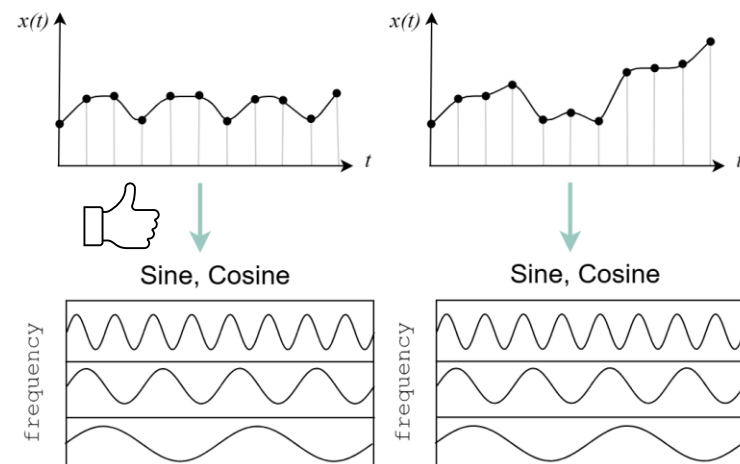
自然言語は定常的な信号ではない場合が多い



非定常的な信号に有効な解析を応用すべきでは？

■ ウェーブレット変換

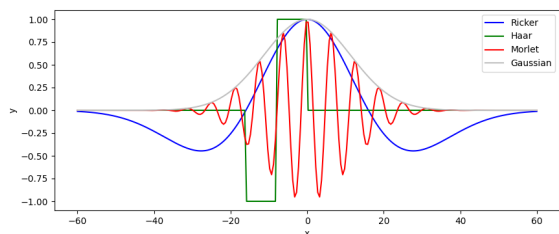
- 時間周波数解析の一種
- 信号を時間領域と周波数領域へと変換
- 非定常的な信号の解析に有効



■ ウェーブレット

- ある範囲にのみ波が局在する波形の総称

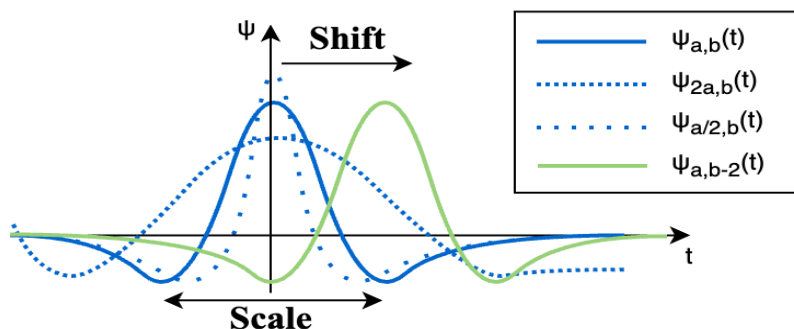
$$\psi_{a,b}(t) = \frac{1}{\sqrt{a}} \psi\left(\frac{t-b}{a}\right).$$



例えば青色のウェーブレットは
下の式で計算される

$$\psi(t) = (1 - t^2) \exp\left(\frac{-t^2}{2}\right).$$

- シフトパラメータとスケールパラメータがある



■ ウェーブレット変換

- ウェーブレットを使った信号解析手法
- ウェーブレット関数と信号の内積を計算することで、どの時点でこういった情報があるのかを解析可能

$$W(a, b) = \sum_{t=0}^{T-1} \psi_{a,b}(t) x(t).$$

(例)スケールaが1、シフトbが0,2のときのウェーブレット変換

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

■ RoPEはウェーブレット変換の一種である

- ・ 位置に対してではなく、ヘッド次元に沿ってウェーブレット変換をしている
- ・ 簡単なウェーブレットと固定のスケールパラメータを使っている

$$\begin{bmatrix} \cos m\theta_1 & -\sin m\theta_1 & 0 & 0 \\ \sin m\theta_1 & \cos m\theta_1 & 0 & 0 \\ 0 & 0 & \cos m\theta_2 & -\sin m\theta_2 \\ 0 & 0 & \sin m\theta_2 & -\cos m\theta_2 \end{bmatrix} \begin{bmatrix} q_0^m \\ q_1^m \\ q_2^m \\ q_3^m \end{bmatrix} \xrightarrow{\text{ヘッド次元}} \begin{bmatrix} W(1, b_0) \\ W(1, b_2) \end{bmatrix} = \begin{bmatrix} \cos \phi_0 & -\sin \phi_1 & 0 & 0 \\ 0 & 0 & \cos \phi_2 & -\sin \phi_3 \end{bmatrix} \begin{bmatrix} x(0) \\ x(1) \\ x(2) \\ x(3) \end{bmatrix} \xrightarrow{\text{時間}}$$

→位置符号化に対してウェーブレット変換を組み込めないか？

■ ALiBiは複数の窓サイズを持つ

- ・ ALiBiがうまくいく理由は複数の大きさの窓サイズでは？

→複数の窓サイズを位置符号化に組み込めないか？

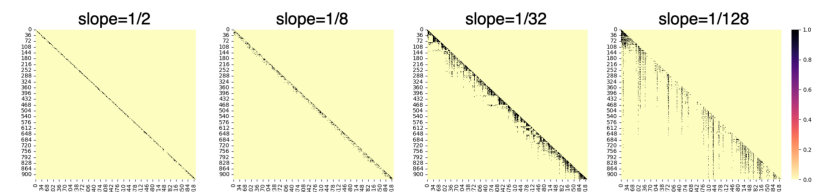


Figure 2: Heatmap of scaled attention scores via softmax normalization in ALiBi without non-overlapping inference. The vertical axis represents the query, while the horizontal axis corresponds to the key in the attention map. For clarity, values of 0.001 or more are mapped to black, while values below that are mapped to yellow. The maximum allowable length of sequences is $L_{\text{train}} = 512$, and the inference length is 1012.

(補足)

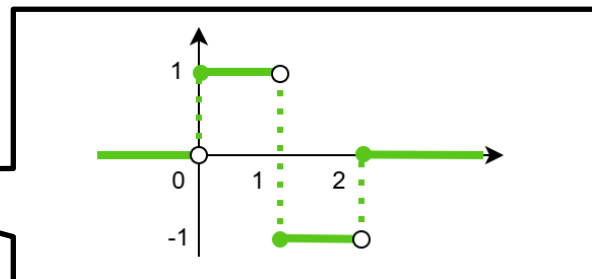
位置符号化とウェーブレット変換の共通点

RoPEはウェーブレット変換の一種である (1/2)

1. 次のハール型のウェーブレット関数を定義

$$\psi(t) = \begin{cases} \cos f(t) & 0 \leq t < 1, \\ -\sin f(t) & 1 \leq t < 2, \\ 0 & \text{otherwise.} \end{cases}$$

(例) ハールウェーブレット



2. スケールを1、シフトを (b_0, b_1) のWTを考える

- $b_j = j - \delta(j)$ と定義し、 $\delta(t)$ は $0 \leq \delta(t) \leq 1$ を満たす単調関数とする
- j が偶数の時、 $\phi_j = f(\delta(j))$ とし、 $\phi_j = \phi_{j+1} = m\theta_{\lceil \frac{j+1}{2} \rceil}$ となるよう f を定義

RoPEの奇数次元と完全に一致する

$$\begin{bmatrix} W(1, b_0) \\ W(1, b_2) \end{bmatrix} = \begin{bmatrix} \cos m\theta_1 & -\sin m\theta_1 & 0 & 0 \\ 0 & 0 & \cos m\theta_2 & -\sin m\theta_2 \end{bmatrix} \begin{bmatrix} x(0) \\ x(1) \\ x(2) \\ x(3) \end{bmatrix}$$

RoPEはウェーブレット変換の一種である (2/2)

RoPE

ウェーブレット変換

$$\begin{bmatrix} \cos m\theta_1 & -\sin m\theta_1 & 0 & 0 \\ \sin m\theta_1 & \cos m\theta_1 & 0 & 0 \\ 0 & 0 & \cos m\theta_2 & -\sin m\theta_2 \\ 0 & 0 & \sin m\theta_2 & -\cos m\theta_2 \end{bmatrix} \begin{bmatrix} q_0^m \\ q_1^m \\ q_2^m \\ q_3^m \end{bmatrix}.$$

$$\begin{bmatrix} W(1, b_0) \\ W(1, b_2) \end{bmatrix} = \begin{bmatrix} \cos m\theta_1 & -\sin m\theta_1 & 0 & 0 \\ 0 & 0 & \cos m\theta_2 & -\sin m\theta_2 \end{bmatrix} \begin{bmatrix} x(0) \\ x(1) \\ x(2) \\ x(3) \end{bmatrix}$$

赤枠と緑枠のところが完全に一致する
→RoPEはウェーブレット変換の一種であることがわかった

ALiBiは複数の窓サイズ(分解能)を持つ

■ ALiBiのヘッド毎のAttention Map

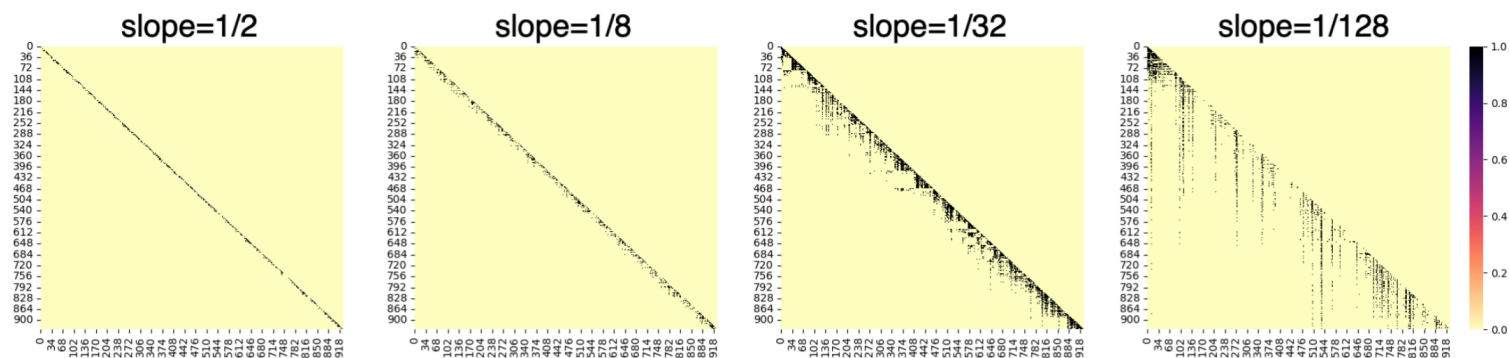


Figure 2: Heatmap of scaled attention scores via softmax normalization in ALiBi without non-overlapping inference. The vertical axis represents the query, while the horizontal axis corresponds to the key in the attention map. For clarity, values of 0.001 or more are mapped to black, while values below that are mapped to yellow. The maximum allowable length of sequences is $L_{\text{train}} = 512$, and the inference length is 1012.

- ヘッド毎にSliding Windowの窓サイズが違う
- 傾き m によって窓サイズが変化している

→ALiBiは**複数窓サイズを持つ**Sliding Windowのような動きをしている

RoPEがウェーブレット変換だとすると次の欠点がある

- ① 時間軸に沿った変換をしていない
- ② スケールaが一定
- ③ 簡単なウェーブレットを使っている

$$\begin{bmatrix} \cos m\theta_1 & -\sin m\theta_1 & 0 & 0 \\ \sin m\theta_1 & \cos m\theta_1 & 0 & 0 \\ 0 & 0 & \cos m\theta_2 & -\sin m\theta_2 \\ 0 & 0 & \sin m\theta_2 & -\cos m\theta_2 \end{bmatrix} \begin{bmatrix} q_0^m \\ q_1^m \\ q_2^m \\ q_3^m \end{bmatrix}.$$

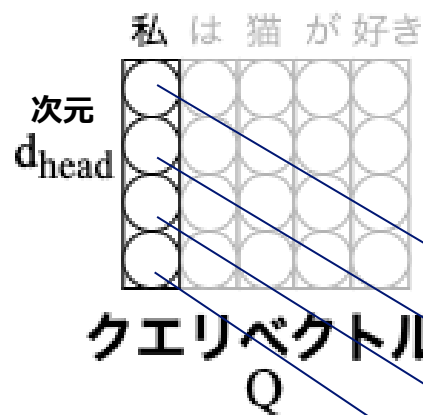
RoPE

$$\begin{bmatrix} W(1, b_0) \\ W(1, b_2) \end{bmatrix} = \begin{bmatrix} \cos \phi_0 & -\sin \phi_1 & 0 & 0 \\ 0 & 0 & \cos \phi_2 & -\sin \phi_3 \end{bmatrix} \begin{bmatrix} x(0) \\ x(1) \\ x(2) \\ x(3) \end{bmatrix}$$

ウェーブレット変換

RoPEがウェーブレット変換だとすると次の欠点がある

① 時間軸に沿った変換をしていない

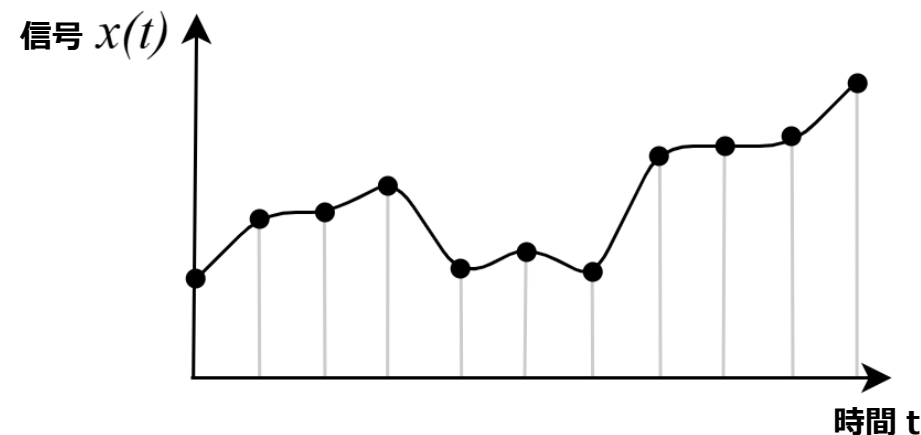


前から $m=0$ 番目の単語
のクエリベクトル

$$\begin{bmatrix} \cos m\theta_1 & -\sin m\theta_1 & 0 & 0 \\ \sin m\theta_1 & \cos m\theta_1 & 0 & 0 \\ 0 & 0 & \cos m\theta_2 & -\sin m\theta_2 \\ 0 & 0 & \sin m\theta_2 & -\cos m\theta_2 \end{bmatrix} \cdot \begin{bmatrix} q_0^m \\ q_1^m \\ q_2^m \\ q_3^m \end{bmatrix}$$

次元

RoPE



$$\begin{bmatrix} W(1, b_0) \\ W(1, b_2) \end{bmatrix} = \begin{bmatrix} \cos \phi_0 & -\sin \phi_1 & 0 & 0 \\ 0 & 0 & \cos \phi_2 & -\sin \phi_3 \end{bmatrix} \begin{bmatrix} x(0) \\ x(1) \\ x(2) \\ x(3) \end{bmatrix}$$

時間

ウェーブレット変換

RoPEがウェーブレット変換だとすると次の欠点がある

② スケールaが一定

- ALiBiの分析から複数の分解能は位置符号化にも有効
- ウェーブレット変換も複数のスケール(分解能)を設定できる

この範囲しか波(値の変化)がない

$$\begin{bmatrix} \cos m\theta_1 & -\sin m\theta_1 & 0 & 0 \\ \sin m\theta_1 & \cos m\theta_1 & 0 & 0 \\ 0 & 0 & \cos m\theta_2 & -\sin m\theta_2 \\ 0 & 0 & \sin m\theta_2 & -\cos m\theta_2 \end{bmatrix} \begin{bmatrix} q_0^m \\ q_1^m \\ q_2^m \\ q_3^m \end{bmatrix}.$$

RoPE

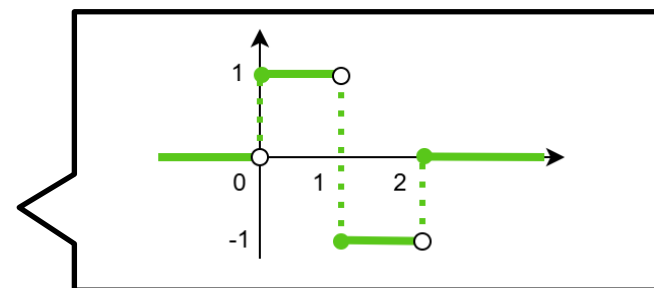
RoPEがウェーブレット変換だとすると次の欠点がある

③ 簡単なウェーブレットを使っている

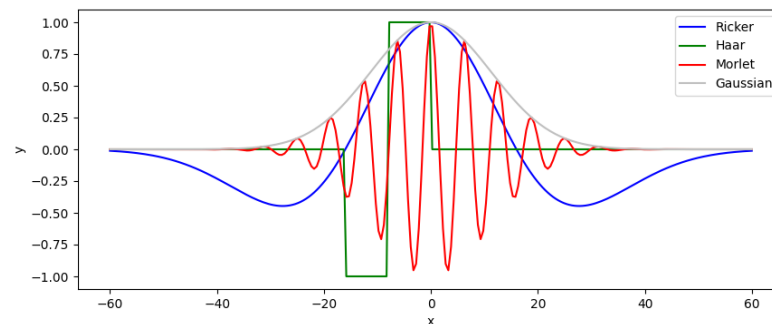
- RoPEだとハール型のウェーブレット

$$\psi(t) = \begin{cases} \cos f(t) & 0 \leq t < 1, \\ -\sin f(t) & 1 \leq t < 2, \\ 0 & \text{otherwise.} \end{cases}$$

(例) ハールウェーブレット



- さらに複雑な形のウェーブレットのほうが有効である可能性



RoPEがウェーブレット変換だとすると次の欠点がある

- ① 時間軸に沿った変換をしていない
- ② スケール a が一定
- ③ 簡単なウェーブレットを使っている

ALiBiの分析から

複数の窓サイズは外挿に有効である可能性

-> RoPEの欠点を解決し、ALiBiの良さを組みこんだ位置符号化を考えたい

提案手法

RPEの学習可能埋め込みをウェーブレット関数に置き換えてみる

- RPE (Relative Position Representation) [7]

- 学習可能埋め込みで表現された相対位置とクエリの内積を計算し、注意スコアに加算

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

	私	は	猫	が	好き
私	0	1			
は	-1	0	1		
猫		-1	0	1	
が			-1	0	1
好き				-1	0
相対位置					

- 決められた個数だけ相対位置を表現する、clippingをする
 - (例) 前後 16 トークンのみ相対位置を表現

提案手法：Wavelet-based Positional Representation

RPEの学習可能埋め込みをウェーブレット関数に置き換えてみる

■ Wavelet-based Positional Representation

- ① 相対位置 $p_{m,n}$ を学習可能埋め込みからウェーブレット関数に変更

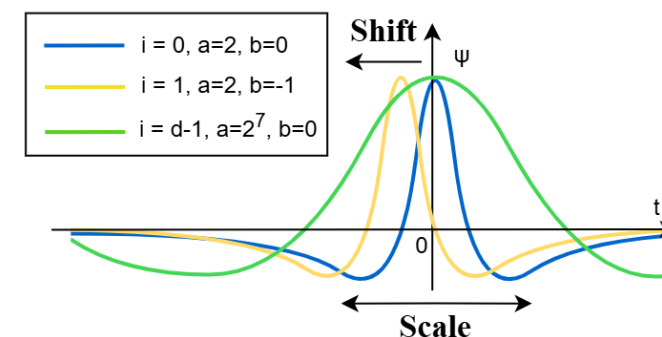
$$e_{m,n} = \frac{q_m k_n^T + q_m (\boxed{p_{m,n}})^T}{\sqrt{d}}, \quad \psi(t) = (1 - t^2) \exp\left(\frac{-t^2}{2}\right).$$

- ② ウェーブレット関数は複数のスケール a ・シフト b を組み合わせる

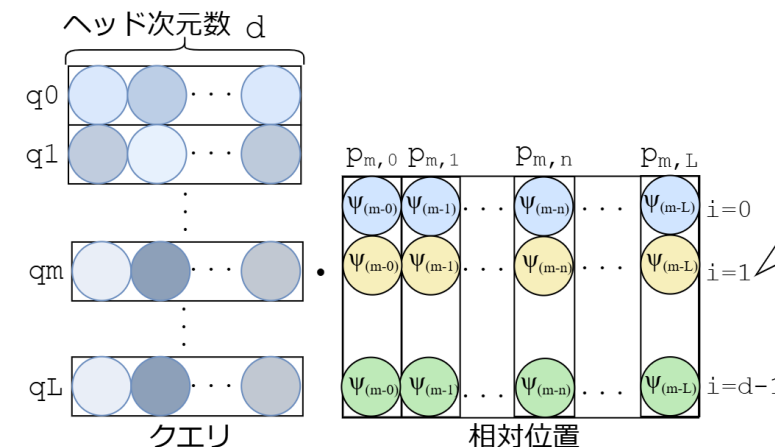
$$(a, b) \in \{2^0, 2^1, 2^2, \dots, 2^{s-1}\} \times \{0, 1, 2, 3, \dots, \frac{d}{s} - 1\}.$$

- ③ clippingを行わず全範囲の位置を計算

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



ウェーブレット関数 $\Psi(t)$



提案手法：Wavelet-based Positional Representation

RPEの学習可能埋め込みをウェーブレット関数に置き換えてみる

■ Wavelet-based Positional Representation

- ① 相対位置 $p_{m,n}$ を学習可能埋め込みからウェーブレット関数に変更

$$e_{m,n} = \frac{q_m k_n^T + q_m (p_{m,n})^T}{\sqrt{d}}, \quad \psi(t) = (1 - t^2) \exp\left(\frac{-t^2}{2}\right)$$

ウェーブレットをヘッド次元方向ではなく、位置方向に適用できる

簡単なハールではなく、複雑なリッカーウェーブレットを使う

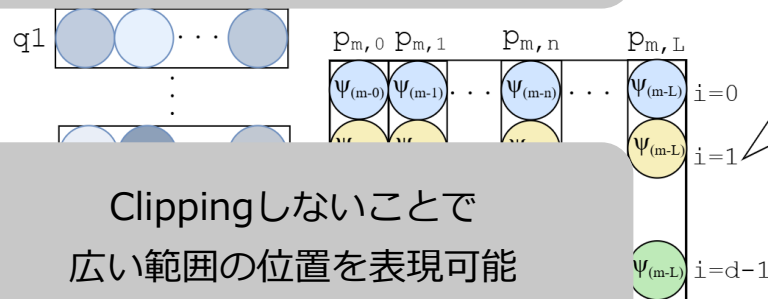
- ② ウェーブレット関数は複数のスケール a ・シフト b を組み合わせる

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

ウェーブレット関数 $\psi(a)$
複数のスケールを採用することで
広い範囲から狭い範囲の相対位置を表現できる

- ③ clippingを行わず全範囲の位置を計算

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



Clippingしないことで
広い範囲の位置を表現可能

スケールパラメータの選び方

波が局在する範囲が最大学習長以下になるよう選択

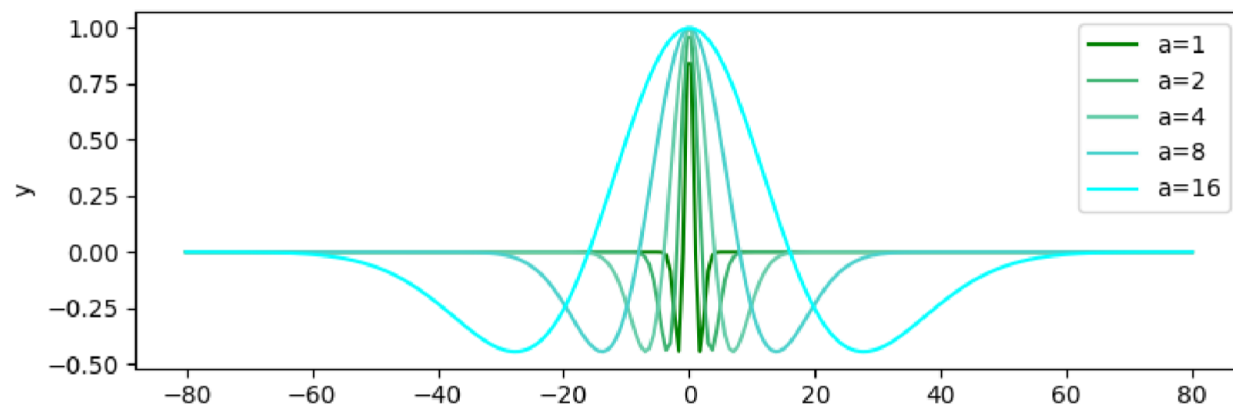


Figure 4: Graph of compared Ricker wavelet functions with $a = [2^0, 2^1, 2^2, 2^3, 2^4]$

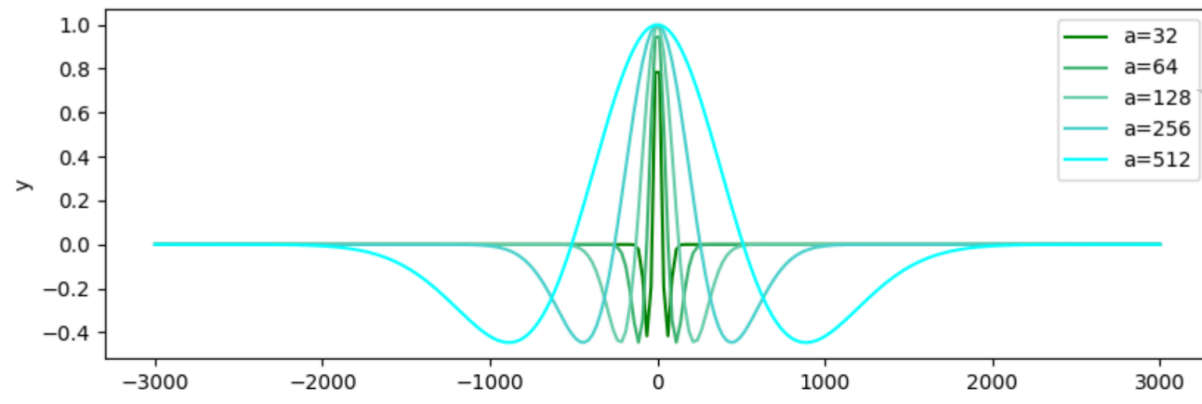


Figure 5: Graph of compared Ricker wavelet functions with $a = [2^5, 2^6, 2^7, 2^8, 2^9]$

例えば、学習時の最大長が512のときはa=128が512までに収束するように見えたのでa=128を選択

- Transformerモデル
 - 埋め込み次元1024,ヘッド数8
 - ヘッド次元128,レイヤー数16
 - 実装はfairseqベース
- データセット：wikitext-103
- 評価：Perplexity
- 比較手法
 - NoPE：位置符号化を一切使わない
 - RoPE： $\Theta=10000$ と $\Theta=500000$ の場合を検証
 - Transformer-XL：Transformer-XLで採用された位置符号化、RPEの学習可能埋め込みを正弦波に変更したもの
 - ALiBi：原論文と同じ設定

* Llama-2-7Bモデルでも簡単な事前学習と比較検証を行っています。
本スライドの付録に結果を記載しています。

- Wikitext103データセットのPerplexity(PPL)で評価
 - どのような条件/長さにおいても最も提案手法のPPLが低い
 - RoPEはSliding Windowを使うか使わないかで外挿したときの性能が大きく異なる
 - NoPEは長系列では有効ではなさそう

Table 1: Perplexity of validation set in extrapolation experiments using Wikitext-103. Maximum allowable lengths of sequences in pre-training are $L_{\text{train}} = 512$ and $L_{\text{train}} = 1024$.

		Sequence Length									
		$L_{\text{train}} = 512$						$L_{\text{train}} = 1024$			
	pos	128	256	512	1012	1512	2512	1024	1524	3024	5024
Perplexity in Non-overlapping Inference with L_{train}											
NoPE(Kazemnejad et al., 2023)	-	26.38	23.23	21.53	21.52	21.53	21.53	20.81	21.52	21.49	21.45
RoPE (Su et al., 2021)	abs	23.82	20.98	19.39	19.35	19.39	19.38	18.42	19.51	19.52	19.48
RoPE (Xiong et al., 2024)	abs	23.81	20.95	19.35	19.32	19.35	19.33	18.50	19.53	19.54	19.50
Trans-XL (Dai et al., 2019)	rel	24.16	21.53	19.96	19.92	19.93	19.96	18.67	19.75	19.74	19.70
ALiBi(Press et al., 2022)	rel	24.18	21.32	19.69	19.64	19.69	19.64	18.66	19.64	19.65	19.62
Wavelet(Ricker)	rel	23.64	20.82	19.19	19.15	19.17	19.20	18.26	19.30	19.34	19.26
Perplexity without Non-overlapping Inference											
NoPE(Kazemnejad et al., 2023)	-	26.38	23.23	21.53	21.03	21.58	48.48	20.81	20.45	22.11	59.37
RoPE (Su et al., 2021)	abs	23.82	20.98	19.39	23.25	44.38	93.94	18.42	18.29	33.20	122.52
RoPE (Xiong et al., 2024)	abs	23.81	20.95	19.35	23.70	40.39	77.90	18.50	18.30	29.25	83.43
Trans-XL(Dai et al., 2019)	rel	24.16	21.53	19.96	19.09	18.92	19.05	18.67	18.25	18.17	18.76
ALiBi(Press et al., 2022)	rel	24.18	21.32	19.69	18.71	18.42	18.41	18.66	18.14	17.86	17.88
Wavelet(Ricker)	rel	23.64	20.82	19.19	18.23	18.00	17.99	18.26	17.13	17.14	17.44

Sliding window(S=512)
を使ってPPLを計算

Sliding windowを使わず
PPLを計算

外挿→

- 学習時の最大長512, 推論時の長さ1024で外挿したときのAttention Map

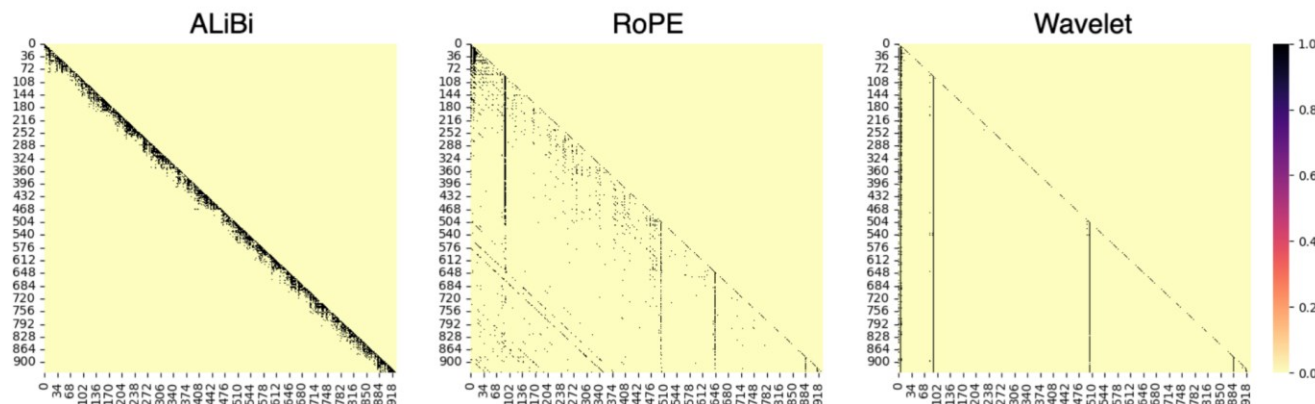


Figure 3: Heatmap of scaled attention scores via softmax normalization in 4th head after softmax operation without non-overlapping inference. The vertical axis represents the query, while the horizontal axis corresponds to the key. For clarity, values of 0.001 or more are mapped to black, while values below that are mapped to yellow. The maximum allowable length of sequences in pre-training is $L_{\text{train}} = 512$ and the inference length is 1012. See Appendix [A.12](#) for other heads.

- 提案手法は受容野を制限することなく生成・外挿ができています
- RoPEは外挿の範囲になると対角線が出現→この対角線は外挿できていない特徴である可能性

おわりに

裏のお気持ち

- 正弦波に集中してる位置符号化に疑問をずっと持っていた
- フーリエ変換のような特徴がうまくいくとしても、自然言語にはマッチしないと思っていた
- 3年前のある日、RPEをウェーブレット関数に置き換えたらうまくいったのが始まり
 - 別のテーマをやることになっていたので3年前そのままお蔵入りに
 - 異動をきっかけに位置符号化に再着手できた
 - 共著者の長谷川さん、西田さん、齊藤さん本当にありがとうございました
- 位置符号化にはまだ謎がいっぱい
 - 相対位置取れてるっていうけどほんまか？0の選び方とか適当じゃない？
 - もっとも良い位置符号化はありそう

	私	は	猫	が	好き
私	0	1			
は	-1	0	1		
猫		-1	0	1	
が			-1	0	1
好き				-1	0

相対位置

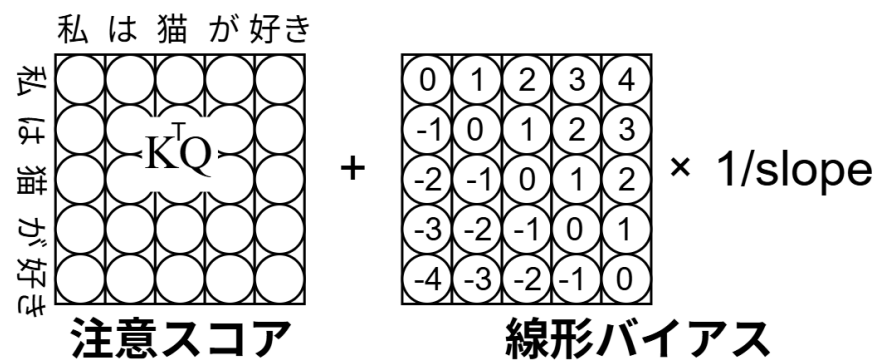
Appendix

提案手法補足

相対位置を使った位置符号化

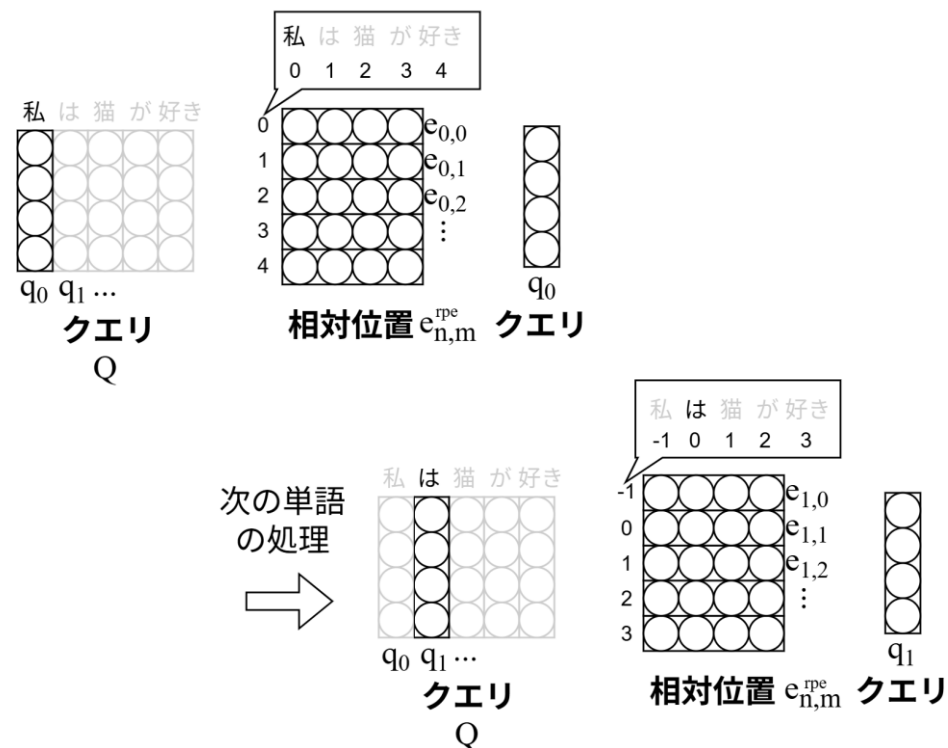
ALiBi [Press+, 2021]

- 現在の単語の後の位置を全て表現



RPE [Shaw+, 2019]

- 現在の単語の前後の位置を表現
- 決められた個数だけ、学習可能なパラメータで表現する



位置符号化まとめ

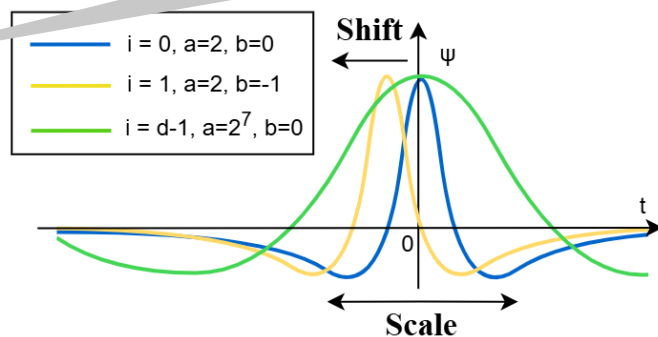
	RoPE	ALiBi	RPE	本技術																																																																																
位置表現	私 は 猫 が 好き <table><tr><td>0</td><td>1</td><td>2</td><td>3</td><td>4</td></tr></table> 絶対位置	0	1	2	3	4	私 は 猫 が 好き <table><tr><td>0</td><td></td><td></td><td></td><td></td></tr><tr><td>-1</td><td>0</td><td></td><td></td><td></td></tr><tr><td>-2</td><td>-1</td><td>0</td><td></td><td></td></tr><tr><td>-3</td><td>-2</td><td>-1</td><td>0</td><td></td></tr><tr><td>-4</td><td>-3</td><td>-2</td><td>-1</td><td>0</td></tr></table>	0					-1	0				-2	-1	0			-3	-2	-1	0		-4	-3	-2	-1	0	私 は 猫 が 好き <table><tr><td>0</td><td>1</td><td></td><td></td><td></td></tr><tr><td>-1</td><td>0</td><td>1</td><td></td><td></td></tr><tr><td></td><td>-1</td><td>0</td><td>1</td><td></td></tr><tr><td></td><td></td><td>-1</td><td>0</td><td>1</td></tr><tr><td></td><td></td><td></td><td>-1</td><td>0</td></tr></table> 相対位置	0	1				-1	0	1				-1	0	1				-1	0	1				-1	0	私 は 猫 が 好き <table><tr><td>0</td><td></td><td></td><td></td><td></td></tr><tr><td>-1</td><td>0</td><td></td><td></td><td></td></tr><tr><td>-2</td><td>-1</td><td>0</td><td></td><td></td></tr><tr><td>-3</td><td>-2</td><td>-1</td><td>0</td><td></td></tr><tr><td>-4</td><td>-3</td><td>-2</td><td>-1</td><td>0</td></tr></table>	0					-1	0				-2	-1	0			-3	-2	-1	0		-4	-3	-2	-1	0
0	1	2	3	4																																																																																
0																																																																																				
-1	0																																																																																			
-2	-1	0																																																																																		
-3	-2	-1	0																																																																																	
-4	-3	-2	-1	0																																																																																
0	1																																																																																			
-1	0	1																																																																																		
	-1	0	1																																																																																	
		-1	0	1																																																																																
			-1	0																																																																																
0																																																																																				
-1	0																																																																																			
-2	-1	0																																																																																		
-3	-2	-1	0																																																																																	
-4	-3	-2	-1	0																																																																																
アプローチ	- クエリと回転行列の内積を計算 - キーと回転行列の内積を計算 私 は 猫 が 好き <table><tr><td>0</td><td>1</td><td>2</td><td>3</td><td>4</td></tr></table> クエリベクトル Q RoPE $\begin{bmatrix} \cos\theta_1 & -\sin\theta_1 \\ \sin\theta_1 & \cos\theta_1 \end{bmatrix} \cdot \begin{bmatrix} 0 \\ 1 \end{bmatrix}$ 私	0	1	2	3	4	クエリとキーの内積を計算した注意スコアに線形的に計算した相対位置を足す 私 は 猫 が 好き 注意スコア 線形バイアス $\begin{bmatrix} 0 & 1 & 2 & 3 & 4 \\ -1 & 0 & 1 & 2 & 3 \\ -2 & -1 & 0 & 1 & 2 \\ -3 & -2 & -1 & 0 & 1 \\ -4 & -3 & -2 & -1 & 0 \end{bmatrix} \times 1/\text{slope}$	学習可能なパラメータで相対位置を表現し、クエリと内積を計算する 私 は 猫 が 好き クエリ Q 相対位置 $e_{n,m}^{\text{rpe}}$ クエリ 次の単語の処理	ウェーブレット関数ですべての相対位置を表現し、クエリと内積を計算する 私 は 猫 が 好き クエリ Q 相対位置 $e_{n,m}^{\text{wpe}}$ クエリ 次の単語の処理																																																																											
0	1	2	3	4																																																																																
特徴	- 計算コストが低い - 外挿ができないので、再学習が必要 - 短い文での性能がよい	- 計算コストが低い - 外挿可能 - 先頭付近の単語の情報が利用できない - 短い文での性能が悪い	- 外挿は不可能 - RoPEより性能が劣る	- 外挿可能 - 短い文でも性能がよい - 先頭付近の単語の情報も利用できる																																																																																

RPEとの比較 - 位置の表現範囲が違う

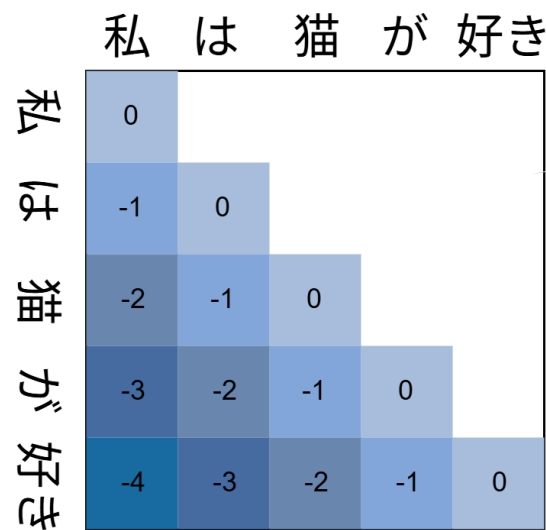
ウェーブレット位置符号化

- 現在の単語の前後の位置を表現
- 全ての単語の位置をウェーブレット関数で表現する
- 複数のスケールとシフトのウェーブレット関数を使う

ウェーブレット関数で
位置を表現する

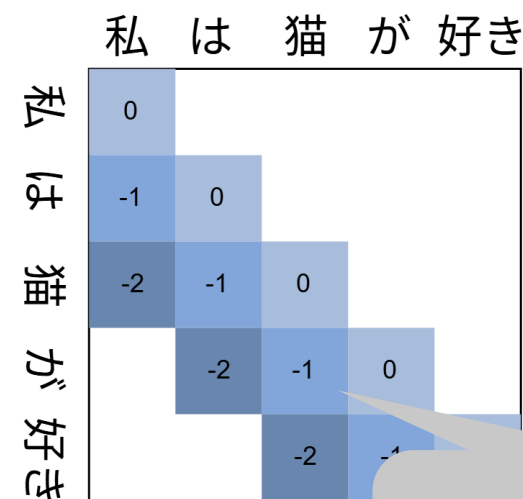


ウェーブレット関数 $\psi(t)$



RPE [Shaw+, 2019]

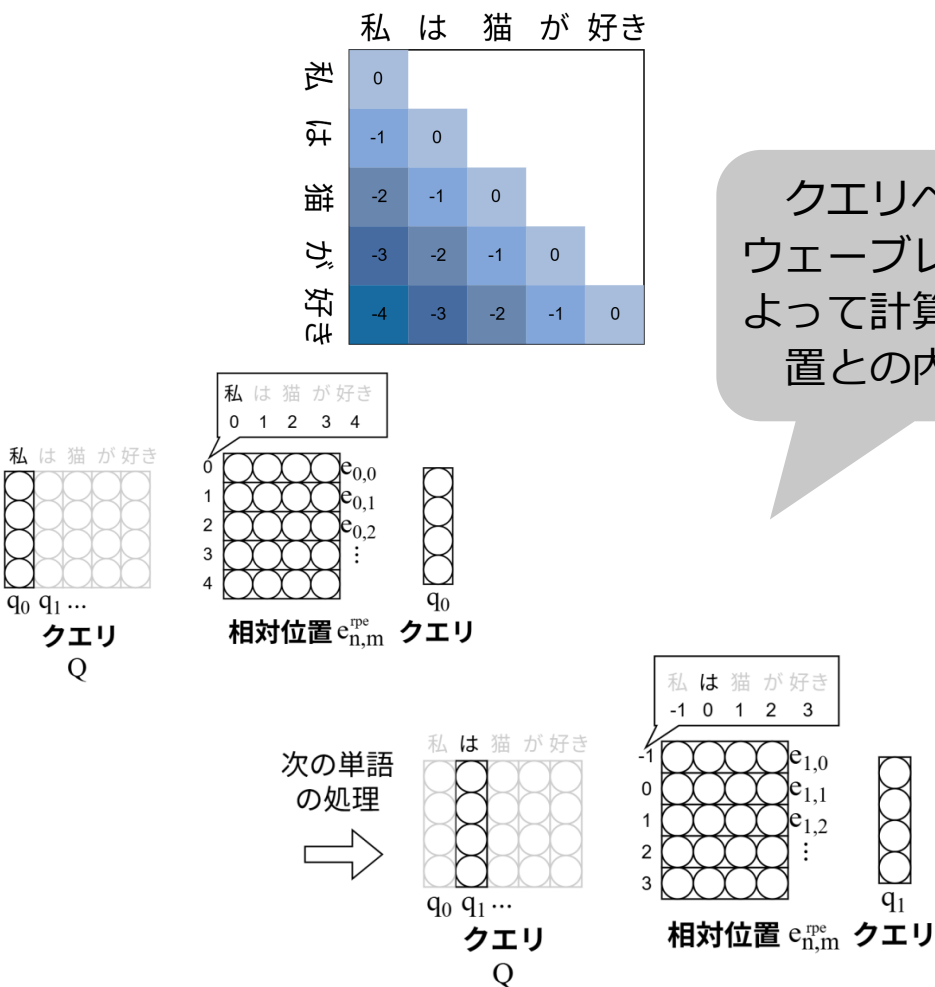
- 現在の単語の前後の位置を表現
- 決められた個数だけ、学習可能なパラメータで表現する



1 個ずつパラメータを学習して位置を表現する

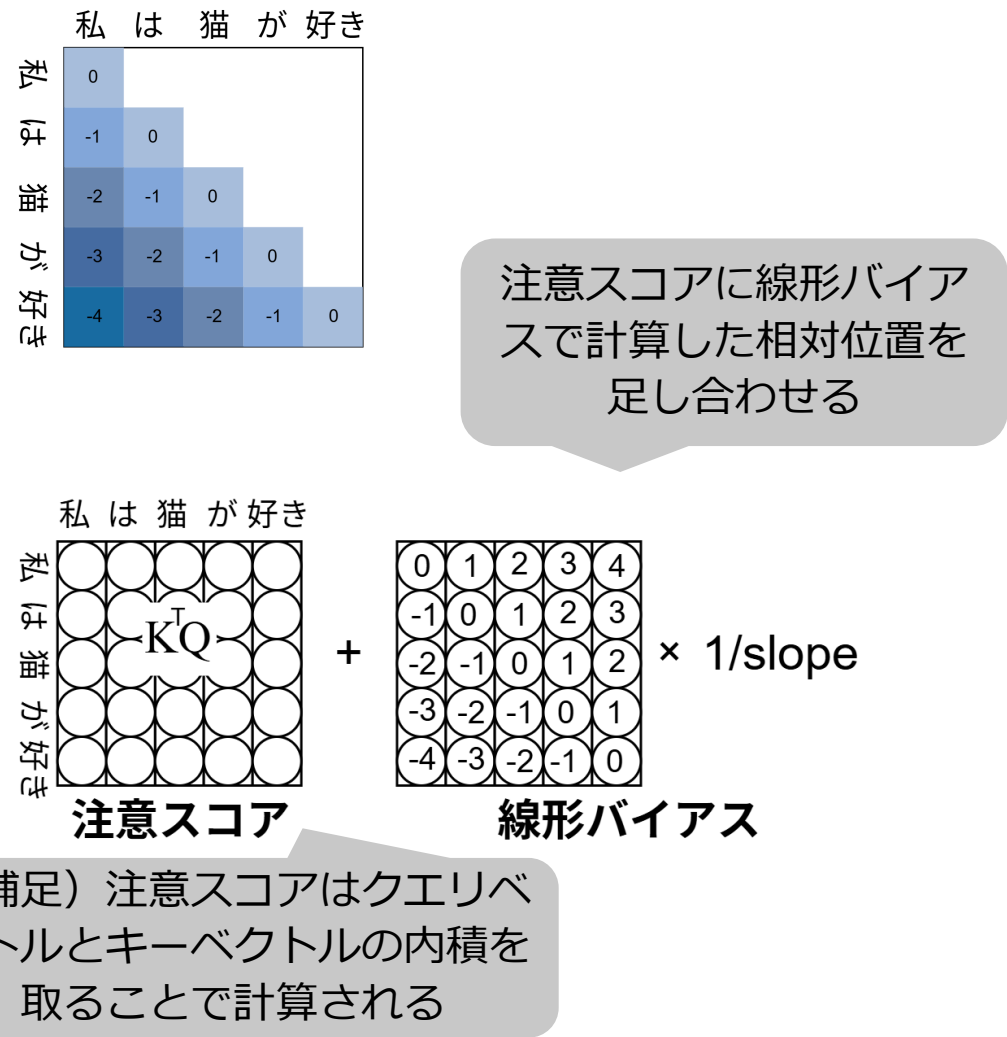
ALiBiとの比較 - モデル内での組み込み方が違う

ウェーブレット位置符号化



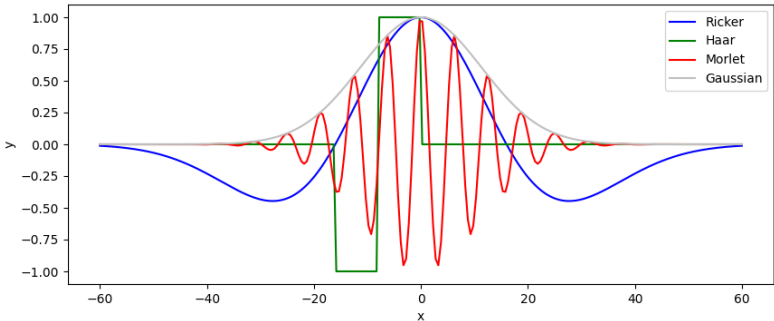
クエリベクトルとウェーブレット関数によって計算した相対位置との内積を取る

ALiBi [Press+, 2021]



実験補足

他のウェーブレット関数での実験



		Sequence Length									
		$L_{\text{train}} = 512$						$L_{\text{train}} = 1024$			
	pos	128	256	512	1012	1512	2512	1024	1524	3024	5024
Perplexity in Non-overlapping Inference with L_{train}											
NoPE (Kazemnejad et al., 2023)	-	26.38	23.23	21.53	21.52	21.53	21.53	20.81	21.52	21.49	21.45
RoPE (Su et al., 2021)	abs	23.82	20.98	19.39	19.35	19.39	19.38	18.42	19.51	19.52	19.48
RoPE (Xiong et al., 2024)	abs	23.81	20.95	19.35	19.32	19.35	19.33	18.50	19.53	19.54	19.50
Trans-XL (Dai et al., 2019)	rel	24.16	21.53	19.96	19.92	19.93	19.96	18.67	19.75	19.74	19.70
ALiBi (Press et al., 2022)	rel	24.18	21.32	19.69	19.64	19.69	19.64	18.66	19.64	19.65	19.62
Wavelet(Ricker)	rel	23.64	20.82	19.19	19.15	19.17	19.20	18.26	19.30	19.34	19.26
Perplexity without Non-overlapping Inference											
NoPE (Kazemnejad et al., 2023)	-	26.38	23.23	21.53	21.03	21.58	48.48	20.81	20.45	22.11	59.37
RoPE (Su et al., 2021)	abs	23.82	20.98	19.39	23.25	44.38	93.94	18.42	18.29	33.20	122.52
RoPE (Xiong et al., 2024)	abs	23.81	20.95	19.35	23.70	40.39	77.90	18.50	18.30	29.25	83.43
Trans-XL (Dai et al., 2019)	rel	24.16	21.53	19.96	19.09	18.92	19.05	18.67	18.25	18.17	18.76
ALiBi (Press et al., 2022)	rel	24.18	21.32	19.69	18.71	18.42	18.41	18.66	18.14	17.86	17.88
Wavelet(Ricker)	rel	23.64	20.82	19.19	18.23	18.00	17.99	18.26	17.13	17.14	17.44
Haar (Fixed scale)	rel	24.98	22.07	20.49	51.61	116.87	299.26	-	-	-	-
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	-	-	-	-

- Gaussianベースのウェーブレットは有望そう
- Haarでスケールを固定すると性能が下がる
(=RoPEの形はウェーブレット変換としてはミスマッチである可能性)
- Morletのように周波数が高い波だと性能が下がる
(=Transformer-XLで採用されている相対位置に正弦波を使うのはミスマッチである可能性)

スケール/シフトのAblation Study

- シフトとスケールを組み合わせたほうがよさそう
- スケールの幅は最大長を少し超えても問題はない
- スケールを固定してシフトを変更すると性能が低くなるので、**スケールの方が重要そう**

Table 3: Perplexity of validation set in extrapolation experiments using Wikitext-103. Maximum allowable length of sequences in pre-training is $L_{\text{train}} = 512$.

			Sequence Length					
	scale a	shift b	128	256	512	1012	1512	2512
Perplexity without Non-overlapping Inference								
Ricker	$\{2^0, 2^1, \dots, 2^7\}$	$\{0, 1, 2, \dots, 15\}$	<u>23.64</u>	<u>20.82</u>	<u>19.19</u>	18.23	18.00	17.99
Ricker	$\{2^1, 2^2, \dots, 2^8\}$	$\{0, 1, 2, \dots, 15\}$	23.77	20.89	19.25	18.23	17.97	18.02
Ricker	$\{2^2, 2^3, \dots, 2^9\}$	$\{0, 1, 2, \dots, 15\}$	23.92	21.03	19.40	18.41	18.14	18.07
Ricker	$\{2^0, 2^1, 2^2, 2^3\}$	$\{0, 1, 2, \dots, 31\}$	23.96	21.13	19.55	18.87	19.40	21.73
Ricker	$\{2^0, 2^1\}$	$\{0, 1, 2, \dots, 63\}$	24.49	21.60	19.95	20.90	32.01	70.80
Ricker	$\{2^0, 2^1, \dots, 2^{15}\}$	$\{0, 1, 2, \dots, 7\}$	23.74	20.88	19.24	<u>18.22</u>	<u>17.96</u>	17.84
Ricker	$\{2^0, 2^1, \dots, 2^{31}\}$	$\{0, 1, 2, 3\}$	23.75	20.86	19.26	18.24	<u>17.96</u>	<u>17.84</u>
Ricker	$\{2^0, 2^1, \dots, 2^{63}\}$	$\{0, 1\}$	23.75	20.88	19.30	18.31	18.04	18.02
Ricker	$\{2^0, 2^1, \dots, 2^{127}\}$	$\{0\}$	23.97	21.10	19.46	18.50	18.27	18.29
Ricker	$\{2^7\}$	$\{0, 1, 2, \dots, 127\}$	24.35	21.45	19.80	20.68	20.87	21.31
Gaussian	$\{2^0, 2^1, \dots, 2^7\}$	$\{0, 1, 2, \dots, 15\}$	<u>23.77</u>	<u>20.90</u>	<u>19.30</u>	<u>18.31</u>	<u>18.02</u>	<u>17.88</u>
Gaussian	$\{2^1, 2^2, \dots, 2^8\}$	$\{0, 1, 2, \dots, 15\}$	23.92	21.02	19.41	18.41	18.15	18.01
Gaussian	$\{2^2, 2^3, \dots, 2^9\}$	$\{0, 1, 2, \dots, 15\}$	23.98	21.09	19.46	18.43	18.13	17.93
Gaussian	$\{2^0, 2^1, 2^2, 2^3\}$	$\{0, 1, 2, \dots, 31\}$	23.83	29.96	19.33	18.43	18.40	18.94
Gaussian	$\{2^0, 2^1\}$	$\{0, 1, 2, \dots, 63\}$	24.28	21.35	19.70	18.96	19.63	23.14
Gaussian	$\{2^0, 2^1, \dots, 2^{15}\}$	$\{0, 1, 2, \dots, 7\}$	<u>23.72</u>	20.86	<u>19.24</u>	<u>18.24</u>	17.95	<u>17.77</u>
Gaussian	$\{2^0, 2^1, \dots, 2^{31}\}$	$\{0, 1, 2, 3\}$	23.78	<u>20.92</u>	19.29	18.30	<u>18.01</u>	17.85
Gaussian	$\{2^0, 2^1, \dots, 2^{63}\}$	$\{0, 1\}$	23.86	20.98	19.37	18.46	18.20	18.10
Gaussian	$\{2^0, 2^1, \dots, 2^{127}\}$	$\{0\}$	24.21	21.31	19.68	18.71	18.45	18.45
Gaussian	$\{2^7\}$	$\{0, 1, 2, \dots, 127\}$	24.48	21.62	20.05	19.53	22.63	35.23
Haar	-	-	<u>24.98</u>	<u>22.07</u>	<u>20.49</u>	<u>51.61</u>	<u>116.87</u>	<u>299.26</u>
Haar	$\{2^0, 2^1, \dots, 2^7\}$	$\{0, 1, 2, \dots, 15\}$	23.73	20.89	19.27	18.34	18.11	18.17
Morlet	$\{2^0, 2^1, \dots, 2^7\}$	$\{0, 1, 2, \dots, 15\}$	24.15	21.28	19.65	19.02	20.46	26.56

Attention Mapの追加分析

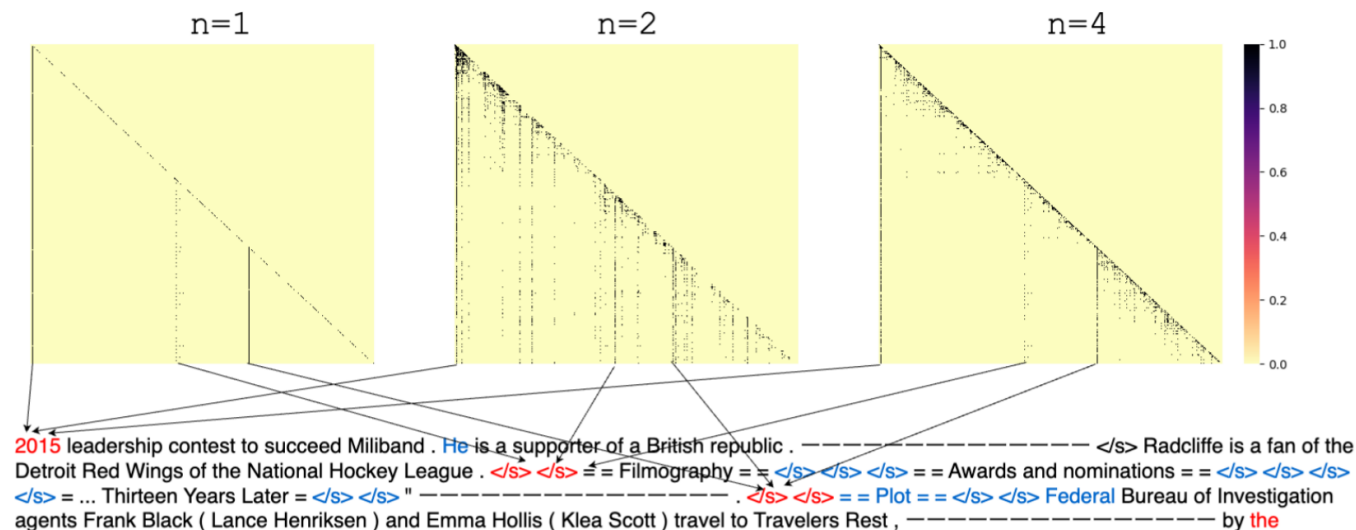


Figure 7: Heatmap of attention score e_{ij} after softmax operation for the proposed method. The maximum sequence length is $L_{max} = 512$, and the sequence length at inference is $L = 1012$. From left to right, $n = 1, 2, 4$ th heads are shown. Scores above 0.01 are mapped in black and the rest in yellow. Words that were always given attention in all heads are shown in red, and words that were frequently given attention only in the $n = 2$ nd head are shown in blue. Sentences are omitted in the middle because they are long with 1012 tokens.

- 先頭トークンを注視可能
- <s>や主語にも注視できている

→この特徴が言語モデルにとって良いかどうかは今後分析予定

- RoPEベースでスケールを複数用意してみた

$$\begin{bmatrix} q_0^m \\ q_1^m \\ q_2^m \\ q_3^m \\ \vdots \\ q_{d-1}^m \\ q_{d-2}^m \end{bmatrix} = \begin{bmatrix} \cos m\theta_1 & -\sin m\theta_1 & 0 & 0 & \dots & 0 & 0 \\ \sin m\theta_1 & \cos m\theta_1 & 0 & 0 & \dots & 0 & 0 \\ \cos m\theta_2 & -\sin m\theta_2 & \cos m\theta_2 & -\sin m\theta_2 & \dots & 0 & 0 \\ \sin m\theta_2 & \cos m\theta_2 & \sin m\theta_2 & \sin m\theta_2 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \cos m\theta_{d/2} & -\sin m\theta_{d/2} & \cos m\theta_{d/2} & -\sin m\theta_{d/2} & \dots & \cos m\theta_{d/2} & -\sin m\theta_{d/2} \\ \sin m\theta_{d/2} & \cos m\theta_{d/2} & \sin m\theta_{d/2} & \sin m\theta_{d/2} & \dots & \sin m\theta_{d/2} & \cos m\theta_{d/2} \end{bmatrix} \begin{bmatrix} q_0^m \\ q_1^m \\ q_2^m \\ q_3^m \\ \vdots \\ q_{d-2}^m \\ q_{d-1}^m \end{bmatrix} \quad (28)$$

- (結果) うまくいかない
 - スケールの範囲の最大値がヘッド次元の大きさしか用意できない
 - メモリ効率、計算効率がRPEより悪い
 - 絶対位置を使っているのは変わらない

- Llama-2と同じ設定、 Θ の値を500000に変更したモデルと比較

- Perplexityの結果

- 長系列のためSliding Window Perplexityで比較
- どのような長さでもRoPEより低い

Table 2: Perplexity in Non-overlapping Inference with $L_{\text{train}} = 4096$.

	Sequence Length			
	4 k	8 k	16 k	32 k
RoPE (Xiong et al., 2024)	9.45	9.33	9.12	8.90
Wavelet	9.00	9.01	8.83	8.60

- LongBenchでの検証

- QAや要約タスクを含んだベンチマーク
- 半分以上のタスクで有効性を確認

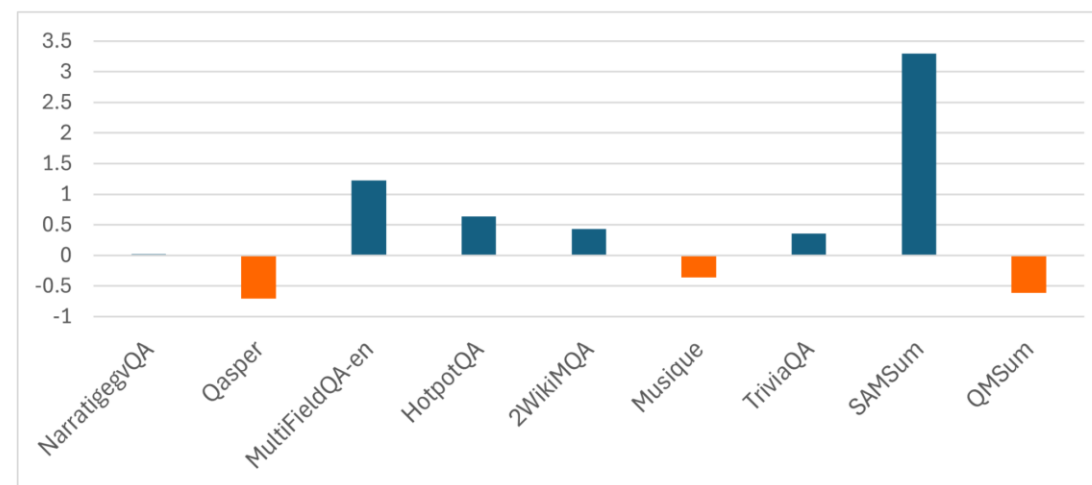


Figure 11: Evaluation results using LongBench(Bai et al., 2024). We evaluated the model pretrained in Section 7. The scores of the difference between the model using the proposed method, which uses wavelet-based position representation, and the model using RoPE are shown. The tasks were evaluated using the dataset, which contains relatively long sentences.