ウェーブレット位置符号化

岡佑依　長谷川拓　西田京介　齋藤邦子
日本電信電話株式会社　NTT 人間情報研究所
yui.oka@ntt.com

概要

本研究では，回転行列による位置符号化 RoPE がウェーブレット変換（WT）の一種として解釈できることを示す．しかし，RoPE は WT の利点を十分に活用できていない．そこで本研究では，正弦波のみに集中していた位置符号化の理論的基盤を拡張し，WT の特性を活かした理論的基盤と新しい位置符号化手法を提案する．提案手法は，従来の手法では難しかった最大系列長を超えた符号化を可能にすることを示す．

1　はじめに

Transformer に基づいた大規模言語モデルは様々な生成タスクにおいて優れた能力を発揮している [1, 2]．しかし，事前学習時の計算資源の制約から，入力文の最大系列長（本稿では，L_{max} と定義する）を事前に決める必要があり，その結果，L_{max} よりも長い文を外挿すると性能が大きく下がる．これはモデルが事前学習時に L_{max} を超える位置の表現を学習していないことが原因である [3]．

位置の表現の学習には，正弦波位置符号化（SPE）[1] や回転行列による位置符号化（RoPE）[4] など，正弦波のような無限かつ周期性を持つものが有効とされている．特に，RoPE は長文を扱う多くの大規模言語モデルで多く採用されている．しかしながら，RoPE の外挿性能は低いため，一般的には位置補間手法 [5] が適用されるが，これらは事前学習に加えて追加の微調整が必要であり，学習コストがさらにかかってしまう．これに対し，線形バイアスを使った位置符号化（ALiBi） [3] は微調整なしで外挿が可能である．しかし，ALiBi は，スライディングウィンドウ [6] のようにアテンションの受容野を制限するため [7]，遠い依存関係にある情報は取得できない課題がある．

本研究では，外挿可能かつアテンションの受容野を制限しない位置符号化について検討する．初

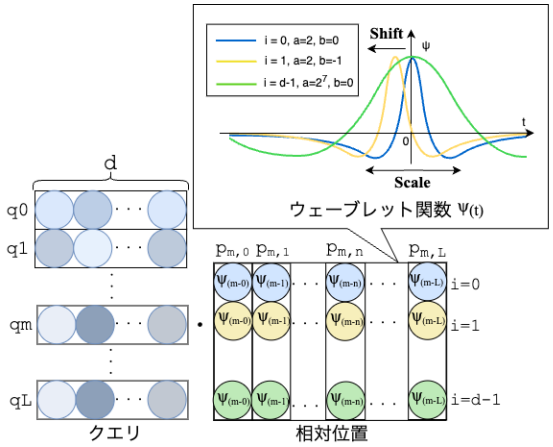

図 1: 提案手法の概要．RPE[8] における学習可能な埋め込みの代わりに，ウェーブレット関数 $\psi_{a,b}$ に基づいて相対位置 $(p_{m,n})^T$ を計算する．この例ではリッカーウェーブレットを示している．スケール a とシフト b は，ヘッドの次元数 d によって変わる．

めに，**RoPE は時間周波数解析手法であるウェーブレット変換（WT）[9] の一形態であると解釈できることを示す．**時間周波数解析は信号を時間と周波数に解析し，時間経過に伴う信号の動的な変化を捉えることができる．文内の各トークンの位置を時間として解釈したとき，RoPE は位置の順序に従って変換を行うのではなく，次元の数に従って変換を行うため，時間経過に伴う信号の動的な変化を捉えることができない．さらに，WT におけるスケールパラメタに対応する窓サイズは一定であり，RoPE は信号を複数のスケールで分析できるという WT の重要な特性を十分に活用できていない．

これらの分析から，本研究では WT を位置符号化に適用した**ウェーブレット位置符号化**を提案する．位置の順序に沿って WT を実行し複数の窓サイズを導入することで，本来の機能を位置符号化に適用する．さらに，相対位置表現（RPE）の方法論 [8] に従うことで，提案手法は比較的容易に実装することができる．外挿機能の実験結果から，従来の位置符号化と比べて提案手法が有効であることがわかった．

2 背景

位置符号化 位置符号化は文中における各トークンの位置を表現し，文の先頭からの位置を表現する絶対位置と，文中における各トークンの相対的な位置を表現する相対位置がある．RoPE [4] は絶対位置の一種であり，回転行列を使用して位置を計算し，クエリとキーに乗算することで位置を表現する．RPE [8] は最大 32 トークンの距離の位置を表す学習可能な埋め込みを使って相対位置を表現する．このような相対位置は系列長に依存しない位置表現であるため，外挿に有効である．特に，ALiBi [3] は，外挿に有効な位置符号化の一つであり，各ヘッドの注意スコアに線形バイアスを加えることで，すべてのトークンの相対位置を表現する．

時間周波数解析 周波数解析 [10] とは，信号や波形を周波数成分に分解し，その特性を調べる手法である．しかし，周波数解析では特定の周波数が「いつ」発生するのかという時間情報は得られない．これに対し，時間周波数解析 [11] は，信号がどの時点でどの周波数成分を持っているかを同時に分析することを可能にする．中でもウェーブレット変換（WT）[9, 12] は，複数のスケールや解像度で信号を分析し，柔軟かつ効果的な解析が可能である．WT は，高周波数成分に対して高い時間分解能を，低周波数成分に対して高い周波数分解能を適応的に提供できるため，非定常信号の解析に適している．

3 RoPE とウェーブレット変換

3.1 ウェーブレット変換（WT）

ウェーブレット（wave-let）[13] は，特定の時間（空間）に局所化し，その中心から離れるにつれて影響が急激に小さくなる波である．実数体 $\mathbb{R}$ 上で定義された関数 ψ が，正方積分関数の空間である正方積分関数空間 $L^2(R)$ に属し，$\int_{-\infty}^{\infty} |\psi(x)|^2\, dx < \infty$ を満たす場合，式 (1) に示すウェーブレット関数と呼ばれる．

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

ここで，b はシフト，$a > 0$ はスケールパラメタである．スケールは，波が局在する範囲と，波の振幅を同時に変化させる．

連続信号から，一定間隔で T 個の値をサンプリングしたと仮定した時，ウェーブレット変換（WT）[9] は，$\psi_{a,b}(t)$ と信号 $x(t)$ の内積を計算することで，信号周波数領域と時間領域に変換する．

$$W(a,b) = \sum_{t=0}^{T-1} \psi_{a,b}(t)x(t). \tag{2}$$

すなわち，WT は信号をスケール a とシフト b による基底関数が張る空間に射影する．例えば，$T = 4$ 個の離散信号を，スケール $a = 1$，シフト $b = [0, 2]$ に変換する場合，式 (2) は行列式で表現できる．

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

この時，波が局在する範囲外では，ウェーブレット行列 Ψ の各要素の値は 0 になるか，0 に近づく．この範囲は，ウェーブレット関数によって異なる．

3.2 RoPE

簡単化のため，ヘッド次元数 $d = 4$ の時の RoPE を式 (3) に示す．

$$\begin{bmatrix} \cos m\theta_1 & -\sin m\theta_1 & 0 & 0 \\ \sin m\theta_1 & \cos m\theta_1 & 0 & 0 \\ 0 & 0 & \cos m\theta_2 & -\sin m\theta_2 \\ 0 & 0 & \sin m\theta_2 & -\cos m\theta_2 \end{bmatrix} \begin{bmatrix} q_0^m \\ q_1^m \\ q_2^m \\ q_3^m \end{bmatrix}. \tag{3}$$

この時，$q^m \in \mathbb{R}^{1\times d}$ は次元数が d の場合の m 番目のクエリ，$\theta_i = 10000^{-2(i-1)/d}$ である．

3.3 理論的解釈

この節では，RoPE は WT の一種と捉えることができることを示す．初めに，ハールウェーブレット [14] から着想を得た関数 $\psi(t)$ を定義する．

$$\psi(t) = \begin{cases} \cos f(t) & 0 \le t < 1, \\ -\sin f(t) & 1 \le t < 2, \\ 0 & \text{otherwise.} \end{cases} \tag{4}$$

$f : \mathbb{R} \to \mathbb{R}$ は，$\int_{-\infty}^{\infty} |\psi(x)|^2\, dx < \infty$ 及び $\int_{-\infty}^{\infty} \psi(t)\, dt = 0$ を満たす関数とする．$\psi(t)$ の波が局在する範囲は $0 \le t < 2$ であり，それ以外の範囲は 0 となる．

次に，式 (3) で示した RoPE の奇数次元が WT の一種であることを示す．$d = 4$ 個の要素を持つ信号 $x(t)$ に対する WT は，スケール $a = 1$，シフト $b \in [b_0, b_2, .., d_{d-2}] = [b_0, b_2]$ の時，以下のように表

現できる．

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

ここで，$b_j = j - \delta(j)$ と定義する．$\delta(t)$ は，$0 \le t \le d-1$ かつ $0 \le \delta(t) \le 1$ を満たす単調関数である．さらに，ϕ_j について，j が奇数の場合は $\phi_j = f(1+\delta(j))$，偶数の場合は $\phi_j = f(\delta(j))$ とする．$j = 0, 2, 4, \ldots, d-2$ に対して，$\phi_j = \phi_{j+1} = m\theta_{\lceil \frac{j+1}{2} \rceil}$ となるように f を定義した時[1)]，この WT は式 (3) の RoPE の奇数次元の変換行列と同一になる．

さらに，以下のハール型ウェーブレット $\psi'(t)$ を使うと偶数次元の RoPE においても同様に成り立つ．

$$\psi'(t) = \begin{cases} \sin f(t) & 0 \le t < 1, \\ \cos f(t) & 1 \le t < 2, \\ 0 & \text{otherwise}. \end{cases} \tag{5}$$

以上のことから，**RoPE は波が $0 \le t < 2$ の範囲に局在するウェーブレットを使った WT と同様の処理をヘッド次元に対して行っている**，と解釈できる．

4 ウェーブレット位置符号化

ウェーブレット変換は，時間とともに変化する信号の動的な変動を捉えるのに有効であり，周期性にとらわれない自然言語の流動性にも有効であると考えられる．さらに，外挿を行う際には，文脈や情報の変化に柔軟に対応できることが重要である．このため，WT は外挿にも有効な手法であると考えられる．しかし，RoPE を WT とみなすと，以下の点で WT の特性を十分に活用できていない．

P1 WT は時間軸で変換を行うのに対し，RoPE はヘッド次元軸で変換しており，単語の位置情報を直接扱っていない．

P2 RoPE はスケール範囲が一定で，多様な分解能を使っていない．

P3 最も単純でステップ上の変化しかとらえることが出来ないハール型を使用しているため，ノイズの影響を受けやすい．

そこで，WT の特性を位置符号化に活用したウェーブレット位置符号化を提案する．

1) $\psi(t)$ がウェーブレットの許容条件を満たすような $f(t)$ が存在する証明は付録 A に記載．

4.1 方法論

我々は RPE の方法論に基づいて WT を位置符号化に適用する[2)]．RPE は，クエリと相対位置埋め込みの内部積を計算することで位置を表現する．

$$e_{m,n} = \frac{q_m k_n^T + q_m (p_{m,n})^T}{\sqrt{d}}, \tag{6}$$

ここで，q_m は長さ L の文の m 番目のクエリ（$q_m \in \mathbb{R}^{1\times d}, 1 \le m \le L$）であり，$k_n$ は n 番目のキー（$k_n \in \mathbb{R}^{1\times d}, 1 \le n \le L$）である．ここで，$p_{m,n}$ は，m 番目のクエリから n 番目のキーまでの相対位置である．RPE では，$p_{m,n}$ を表現するために，スケールを固定した学習可能な埋め込みを使用している．

単語の位置情報 $p_{m,n}$ を表現するために学習可能な埋め込みを使用する代わりに，d-個のウェーブレットを使用して位置を計算する．このように RPE の方法論に WT を組み込むことで **P1** で指摘した単語の位置情報を使って WT が可能となる．

複数のスケール d-個のウェーブレットをスケールパラメタ a，シフトパラメタ b を組み合わせて表現する．$d = 128$ の時，次の 8 個の a と 16 個の b の組み合わせ $(a, b) \in \{2^0, 2^1, 2^2, \ldots 2^7\} \times \{0, 1, 2, 3, \ldots, 15\}$ を使う[3)]．これによって **P2** で指摘したさまざまなスケールパラメタ (分解能) を活用できるようになる．

ウェーブレットの変更 ウェーブレット関数は，以下のリッカーウェーブレット [15] を使う．

$$\psi(t) = (1 - t^2)\exp\left(\frac{-t^2}{2}\right). \tag{7}$$

我々の方法では RPE のようにクリッピングは行わず，文章の長さ全てに対してウェーブレット関数を使って位置を計算する．また，本来のウェーブレットは式 (1) のように振幅も変わるが，アテンションスコアへの影響を考え振幅を変更しない．これによって **P3** で指摘した複雑なウェーブレットを活用できるようになる．まとめると，提案手法における相対位置 $p_{m,n}$ は次のように計算される．

$$p_{m,n} = \left(1 - \left(\frac{m-n-b}{a}\right)^2\right)\exp\left(-\frac{1}{2}\left(\frac{m-n-b}{a}\right)^2\right). \tag{8}$$

2) RoPE を拡張することも検討したが，スケールパラメタの値が d 以下の値に制限されること，計算コストが増加すること，メモリ使用量が増加すること，相対位置の方が外挿に有効であることを踏まえて，RPE の方法論を採用した．

3) 最もよかったパラメタの組み合わせをここでは報告している．その他のパラメタの結果は付録 D に記載．

表 1: 節 5.1 における実験結果 (PPL)

	系列長		
	256	512	2512
NoPE[18]	23.23	21.53	48.48
RoPE($\theta = 10^4$)[4]	20.98	19.39	93.94
RoPE($\theta = 5 \times 10^5$)	20.95	19.35	77.90
Trans-XL[19]	21.53	19.96	19.05
ALiBi[3]	21.32	19.69	18.41
Wavelet (提案手法)	**20.82**	**19.19**	**17.99**

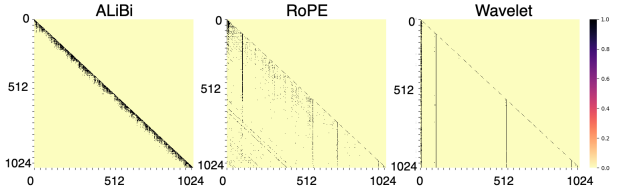

図 2: 系列長が 1024 の時のアテンションスコアのヒートマップ．縦軸がクエリ，横軸がキー．

5 実験

5.1 小規模なモデルでの評価

実験設定 Transformer ベースの小規模な言語モデル [16] を使って比較評価を行なった．学習と評価のデータセットには，WikiText-103[17] を使った．外挿実験で用いたパラメタ設定は [3] と同じものを使った[4]．評価指標には Perplexity（PPL）を用いた．また，学習時の最大入力長は $L_{max} = 512$ とし，系列長が 256，512，2512 のそれぞれで検証した．

ベースライン ベースラインとして，θ の値が 10000 と 500000 に設定した RoPE と比較した．さらに，位置符号化を使わない NoPE[18]，Transformer-XL[19] で採用されている RPE の $p_{m,n}$ を正弦波に変更した位置符号化，そして，ALiBi[3] と比較する．

実験結果 実験結果を表 3 に示す．$L_{max} = 512$ よりも長い文を外挿した場合でも PPL が下がることから，提案手法が外挿性能が向上することがわかった．また，L_{max} よりも短い文についても PPL は下がり，提案した WT が学習済・未学習の両方系列長に対して有効に働いていることを確認できた．さらに，式 (7) 以外の他のウェーブレット関数でも同様の結果であることを確認した[5]．

受容野の制限 図 2 に，アテンションスコアのヒートマップを示す．ALiBi はアテンションの受容野を制限しており，遠い単語の情報を捉えれていない．RoPE は最大入力長 $L_{max} = 512$ を超えると，マップに対角線が現れる．この対角線は外挿ができていない特徴と考えられる．提案手法では，受容野を制限することなく遠い単語の情報を捉えており，対角線のようなものも存在しない．よって，提案手法は受容野を制限することなく外挿可能である．

4） 詳細な実験設定は付録 B に記載．
5） 他のウェーブレット関数の実験結果は付録 C に記載．

表 2: 節 5.2 における実験結果 (PPL)

	系列長			
	4000	8000	16000	32000
RoPE($\theta = 5 \times 10^5$)	9.45	9.33	9.12	8.90
Wavelet (提案手法)	**9.00**	**9.01**	**8.83**	**8.60**

5.2 大規模なモデルでの評価

実験設定 提案手法が大規模なモデルでも有効かどうかを検証する．言語モデル Llama-3-8B[6] と同じモデルの事前学習を行った．学習データセットには Redpajama[20] を利用し，評価には Codeparrot[21] を利用した．評価指標には Perplexity を用いた．また，学習時の最大入力長は $L_{max} = 4098$ とした[4]．

ベースライン Llama-3-8B[6] で採用されている θ の値を 500000 に設定した RoPE と比較する．

実験結果 実験結果を表 2 に示す．結果から，系列長に関わらず提案手法の方が RoPE よりも PPL が低いことがわかった[7]．よって，大規模なモデルにおいても提案手法は有効である可能性がある．

6 おわりに

本研究では，RoPE が WT の一形態として解釈でき，RoPE では WT の特性を活かせていないことを明らかにした．次に，WT の特性が自然言語と外挿に適していると仮定し，WT の特性を活かしたウェーブレット位置符号化を提案した．実験結果から，提案手法は受容野を制限することなく位置の表現が可能であることがわかった．

本研究の最も重要な貢献は，正弦波に集中していた位置符号化の理論的基盤を信号解析手法として確立されている WT に拡張し，新たな研究テーマを開拓したことにある．WT を応用した本手法では外挿が可能となることで追加の再学習が不要となり，計算機資源を多く必要とする最大系列長拡張の学習コスト低減という，産業上重要な課題の解決に貢献できる．

6） https://huggingface.co/meta-llama/Meta-Llama-3-8B
7） 要約や文書 QA などの下流タスクを含む LongBench[22] でも評価を行った実験結果は付録 E に記載．

参考文献

[1] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, L ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In **Advances in Neural Information Processing Systems**, Vol. 30, 2017.

[2] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models. **ArXiv**, Vol. abs/2302.13971, , 2023.

[3] Ofir Press, Noah Smith, and Mike Lewis. Train short, test long: Attention with linear biases enables input length extrapolation. In **International Conference on Learning Representations**, 2022.

[4] Jianlin Su, Yu Lu, Shengfeng Pan, Bo Wen, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding, 2021.

[5] Bowen Peng, Jeffrey Quesnelle, Honglu Fan, and Enrico Shippole. YaRN: Efficient context window extension of large language models. In **The Twelfth International Conference on Learning Representations**, 2024.

[6] Iz Beltagy, Matthew E. Peters, and Arman Cohan. Longformer: The long-document transformer. **arXiv:2004.05150**, 2020.

[7] Ta-Chung Chi, Ting-Han Fan, Alexander Rudnicky, and Peter Ramadge. Dissecting transformer length extrapolation via the lens of receptive field analysis. In **Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 13522–13537, July 2023.

[8] Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. Self-attention with relative position representations. In **Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)**, pp. 464–468. Association for Computational Linguistics, June 2018.

[9] A. Grossmann and J. Morlet. Decomposition of hardy functions into square integrable wavelets of constant shape. **SIAM Journal on Mathematical Analysis**, Vol. 15, No. 4, pp. 723–736, 1984.

[10] Ronald Newbold Bracewell and Ronald N Bracewell. **The Fourier transform and its applications**, Vol. 31999. McGraw-Hill New York, 1986.

[11] Dennis Gabor. Theory of communication. **Journal of the Institution of Electrical Engineers - Part I: General**, Vol. 94, pp. 58–58, 1946.

[12] S.G. Mallat. A theory for multiresolution signal decomposition: the wavelet representation. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, Vol. 11, No. 7, pp. 674–693, 1989.

[13] J. Morlet, G. Arens, E. Fourgeau, and D. Giard. Wave propagation and sampling theory; part i, complex signal and scattering in multilayered media. **Geophysics**, Vol. 47, No. 2, pp. 203–221, 02 1982.

[14] A. Haar. Zur theorie der orthogonalen funktionensysteme. (erste mitteilung). **Mathematische Annalen**, Vol. 69, pp. 331–371, 1910.

[15] Norman Ricker. Wavelet functions and their polynomials. **Geophysics**, Vol. 9, No. 3, pp. 314–323, 07 1944.

[16] Alexei Baevski and Michael Auli. Adaptive input representations for neural language modeling. In **International Conference on Learning Representations**, 2019.

[17] Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. In **International Conference on Learning Representations**, 2017.

[18] Amirhossein Kazemnejad, Inkit Padhi, Karthikeyan Natesan, Payel Das, and Siva Reddy. The impact of positional encoding on length generalization in transformers. In **Thirty-seventh Conference on Neural Information Processing Systems**, 2023.

[19] Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc Le, and Ruslan Salakhutdinov. Transformer-XL: Attentive language models beyond a fixed-length context. In Anna Korhonen, David Traum, and Lluís Màrquez, editors, **Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics**, pp. 2978–2988, Florence, Italy, July 2019. Association for Computational Linguistics.

[20] Maurice Weber, Daniel Y Fu, Quentin Gregory Anthony, Yonatan Oren, Shane Adams, Anton Alexandrov, Xiaozhong Lyu, Huu Nguyen, Xiaozhe Yao, Virginia Adams, Ben Athiwaratkun, Rahul Chalamala, Kezhen Chen, Max Ryabinin, Tri Dao, Percy Liang, Christopher Re, Irina Rish, and Ce Zhang. Redpajama: an open dataset for training large language models. In **The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track**, 2024.

[21] codeparrot. Codeparrot-clean, 2021. https://huggingface.co/datasets/codeparrot/codeparrot-clean/.

[22] Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. LongBench: A bilingual, multitask benchmark for long context understanding. In **Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 3119–3137, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[23] Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. fairseq: A fast, extensible toolkit for sequence modeling. In **Proceedings of NAACL-HLT 2019: Demonstrations**, 2019.

A　ウェーブレットの許容条件

本章では，$\psi(t)$ がウェーブレットの許容条件を満たす関数 f が存在することを示す．ここでは $0 < f(t) \leq 2k\pi (0 \leq t < 2)$ を満たす単調関数 $f(t)$ について考える．ただし，k は $m < 2k\pi$ を満たす最小の自然数とする．

初めに，$0 < f(t) \leq 2k\pi$ を満たす任意の $f(t)$ について，$\int_{-\infty}^{\infty} |\psi(x)|^2 dx < \infty$ を満たすことは自明である．

次にゼロ平均性について考える．$f(t) = 2k\pi t (0 \leq t < 1)$, $f(t) = 2k\pi(t-1) (1 \leq t < 2)$ とした場合，$\theta = f(t)$ とおくと，

$$\int_{-\infty}^{\infty} \psi(t)dt = \int_{0}^{2k\pi} \cos\theta d\theta + \int_{0}^{2k\pi} -\sin\theta d\theta = 0 \quad (9)$$

となり，これはゼロ平均性も満たすため，ψ がウェーブレットとなる f が存在すると言える．

そして，$j = 0, 2, \ldots, d-2$ に対して，$\phi_j (= f(\delta(j))) = \phi_{j+1} (= f(1+\delta(j))) = 2k\pi\delta(t) = m\theta_{\lceil \frac{j+1}{2} \rceil}$ を満たす $\delta(t)$ が存在することも明らかである．つまり $\delta(j) = \frac{m\theta_{\lceil \frac{j+1}{2} \rceil}}{2k\pi} (j = 0, 2, \ldots, d-2)$ を満たす関数を 1 つ選べば良い．

B　実験設定

節 5.1 における実験設定　Transformer ベースの言語モデル [16] を使って比較評価を行なった．データセットには，WikiText-103[17] を使った．WikiText-103 データセットは，1 億 300 万トークン以上の英語版 Wikipedia の記事から構成される．単語埋め込みの次元数 d_{model} は 1024，ヘッド数 n は 8，ヘッドの次元数 d は 128，レイヤー数は 16 である．外挿実験で用いたパラメタ設定は ALiBi の原論文 [3] と同じものを使った．学習エポック数は 205，バッチサイズは 9216 である．学習率は 1.0 とし，学習の過程で 16000step 毎に 1e-7 ずつ更新した．実装には文献 [3] が提供する fairseq[23] ベースのコード[8] を用い，ハイパーパラメタは全てにおいて文献 [3] と同じ設定とした．また，学習時の最大入力長は $L_{max} = 512$ とした．

節 5.2 における実験設定　言語モデル Llama-3-8B[9] と同じモデルの事前学習を行った．学習データセットには Redpajama[20] の 1B トークン分[10] を利用した．単語埋め込みの次元数 d_{model} は 4096，ヘッド数 n は 32，ヘッドの次元数 d は 128，レイヤー数は 32 である．学習エポック数は 1，バッチサイズは 16 である．学習率は 0.0003 とした．実装には huggingface が提供するコード[11] を用いた．また，学習時の最大入力長は $L_{max} = 4096$ とした．

C　ウェーブレット例

図 3 にウェーブレットの例を示す．スケールとシフトは全て同じである．これらのウェーブレットを使った場合の提案手法の性能を評価した．結果から，余弦波を組み込んだウェーブレットであるモルレー以外で外挿性能があり，ガウシアンやリッカーが最も外挿性能が高いことがわかった．

8）https://github.com/ofirpress/attention_with_linear_biases

9）https://huggingface.co/meta-llama/Meta-Llama-3-8B

10）https://huggingface.co/datasets/togethercomputer/RedPajama-Data-1T-Sample

11）https://github.com/huggingface/transformers

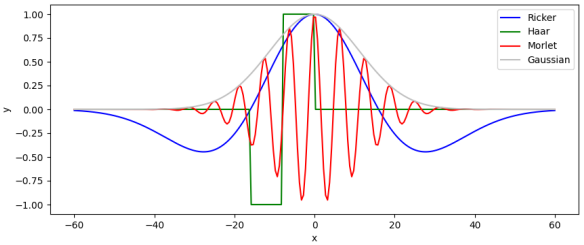

図 3: ウェーブレットの例．青がリッカー，緑がハール，赤がモルレー，灰色がガウシアンである．

表 3: 節 5.1 における実験結果

	系列長		
	256	512	2512
Wavelet(Ricker)	**20.82**	**19.19**	**17.99**
Haar	20.89	19.27	18.17
Morlet	21.28	19.65	26.56
Gaussian	20.90	19.30	17.88

D　Ablation Study

さまざまなシフトパラメタ，スケールパラメタを組み合わせで実験を行った．結果から，スケールパラメタのみ，またはシフトパラメタのみを使う場合よりも組み合わせた方が良い性能を発揮することがわかった．また，$a = \{2^0, 2^1, ..., 2^7\}$ と $b = \{0, 1, 2, ..., 15\}$ の組み合わせが最も安定して性能が良い．

表 4: 実験結果．学習時の最大系列長は $L_{train} = 512$.

	scale a	shift b	系列長 256	512	2512
Ricker	$\{2^0, 2^1, ..., 2^7\}$	$\{0, 1, 2, ..., 15\}$	20.82	19.19	17.99
Ricker	$\{2^1, 2^2 ..., 2^8\}$	$\{0, 1, 2, ..., 15\}$	20.89	19.25	18.02
Ricker	$\{2^2, 2^3 ..., 2^9\}$	$\{0, 1, 2, ..., 15\}$	21.03	19.40	18.07
Ricker	$\{2^0, 2^1, 2^2, 2^3\}$	$\{0, 1, 2, ..., 31\}$	21.13	19.55	21.73
Ricker	$\{2^0, 2^1\}$	$\{0, 1, 2, ..., 63\}$	21.60	19.95	70.80
Ricker	$\{2^0, 2^1 ..., 2^{15}\}$	$\{0, 1, 2, ..., 7\}$	20.88	19.24	17.84
Ricker	$\{2^0, 2^1 ..., 2^{31}\}$	$\{0, 1, 2, 3\}$	20.86	19.26	17.84
Ricker	$\{2^0, 2^1 ..., 2^{63}\}$	$\{0, 1\}$	20.88	19.30	18.02
Ricker	$\{2^0, 2^1 ..., 2^{127}\}$	$\{0\}$	21.10	19.46	18.29
Ricker	$\{2^7\}$	$\{0, 1, 2, ..., 127\}$	21.45	19.80	21.31

E　LongBench での評価

LongBench では正答率と ROUGE スコアで評価を行う．RoPE と提案手法のスコアの差を図 4 に示す．青色は提案手法が RoPE を上回ったスコアを，オレンジ色は RoPE が提案手法を上回ったスコアを示している．

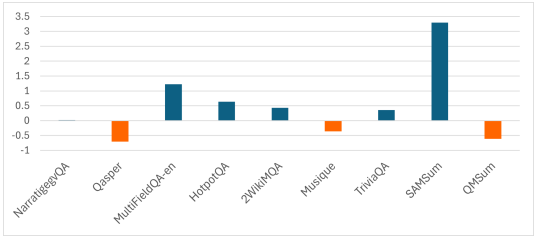

図 4: LongBench における実験結果