自己注意機構における注意の集中が相対位置に依存する仕組み

山本悠士　松崎拓也
東京理科大学 理学部第一部 応用数学科
1419096@ed.tus.ac.jp　matuzaki@rs.tus.ac.jp

概要

自己注意機構において各トークンは自身の周辺のトークンに注意を集中させる傾向がある．これにより，各トークンの出力ベクトルは周辺のトークンからの影響を受け，ニューラル言語モデルは文脈に依存した単語埋め込みを出力できると考えられる．

本研究では，自己注意機構において各トークンが周辺のトークンに注意を集中させるメカニズムを分析する．結果として，自己注意機構は中間層の隠れ状態から位置埋め込みの成分をトークンの位置に関して周期的な波形として抽出し，その波形の位相がずれる方向に注意を集中させていることを示す．

1　はじめに

自然言語処理のタスクを高精度で実行するためには，文脈を捉えることが不可欠である．例えば，回帰型ニューラルネットは時系列順に単語埋め込みを入力することで，畳み込みニューラルネットは周辺の単語埋め込みを集約することで文脈を捉えている．一方，Transformer [1] には単語の位置に依存した処理がモデル内部に存在せず，代わりに単語埋め込みに位置埋め込みを加えたものを入力する．

アテンション重み (式 (4)) を観察することは，自己注意機構の推論過程を解釈する手がかりとなる [2] [3]．このような分析により，自己注意機構において各トークンは自身の周辺に注意を向ける傾向が確認されている [4]．この現象は，自己注意機構が位置に依存しない構造をしているにも関わらず，位置に基づいた推論をしていることを示している．

本研究では，学習に基づく位置埋め込みを用いた自己注意機構に次の性質があることを示す．

- 自己注意機構は，学習された位置埋め込みに存在する周期性を隠れ状態から抽出できる．
- 時系列として見たクエリとキーの間には位相にズレが存在し，注意はこのズレの方向に向く．

以上から，注意が相対位置に依存して集中する現象は，自己注意機構が位置埋め込み由来の周期的な成分を利用することで引き起こされていると言える．

以下，本論文では，2 節で Transformer とその派生モデルの構成について，3 節で注意が相対位置に依存するメカニズムについて述べる．具体的には，まず，注意が相対位置に依存する事実と位置埋め込みの周期性を確認する．次に，自己注意機構内のクエリとキーの共通点と相違点を分析し，それらと注意の方向の関係を示す．最後に，4 節で結論を述べる．

2　Transformer のアーキテクチャ

2.1　位置埋め込み

本節では Transformer，BERT [5] および RoBERTa [6] の位置埋め込みについて説明する．入力文のトークン数を T, 埋め込みを d 次元ベクトルとすると，位置埋め込みは $T \times d$ 行列として定義される．位置埋め込みの各列はトークン位置方向の時系列とみなせることに留意されたい．

Transformer の位置埋め込みは，トークンの位置を pos とするとき，各成分が式 (1-2) で定義される．

$$PE_{(pos,2i)} = \sin(pos/10000^{2i/d}) \tag{1}$$

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

Vaswani らは，式 (1-2) に回転行列を掛けると位置 pos をずらすことができ，この性質が相対位置に基づく注意を学習するように促すと推測した [1]．

$$\begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} \sin x_i \\ \cos x_i \end{bmatrix} = \begin{bmatrix} \sin(x_i+\theta) \\ \cos(x_i+\theta) \end{bmatrix} \tag{3}$$

一方，BERT の位置埋め込みは $T \times d$ 行列のパラメータとして定義され，平均 0，標準偏差 0.02 の正規分布に従う乱数で初期化した状態から学習される．RoBERTa は，BERT とアーキテクチャは同一だが，事前学習の手法が改善されている．特に，BERT は事前学習の 90% を 128 トークンという短い文章

で学習しているのに対して，RoBERTa は常に最大長である 512 トークンの文章で学習しているため，長文に対応することができる．そこで，本研究では各位置での学習回数の偏りがない RoBERTa を用いた．

2.2 マルチヘッド自己注意機構

RoBERTa の各層にあるマルチヘッド自己注意機構は，各層のヘッド数を n とするとき，第 l 層の隠れ状態 $X_l \in \mathbb{R}^{T\times d}$ とパラメータ行列 W^Q_{lh}, W^K_{lh}, $W^V_{lh} \in \mathbb{R}^{d\times(d/n)}$ $(h = 1, \ldots, n)$, $W^O_l \in \mathbb{R}^{d\times d}$ により次のように定義される．以降，第 l 層の h 番目のヘッドを head$(l,\ h)$ と略記する．

$$A_{lh} = \mathrm{softmax}\left(\frac{X_l W^Q_{lh}(X_l W^K_{lh})^T}{\sqrt{d/n}}\right) \tag{4}$$

$$V_{lh} = A_{lh} X_l W^V_{lh} \tag{5}$$

$$\mathrm{MultiHead}_l = \mathrm{concat}(V_{l1}, \ldots, V_{ln}) W^O_l \tag{6}$$

以下，すべての層，ヘッドで共通する議論では添字の l, h を省略する．行列 XW^Q, XW^K, A は，それぞれクエリ，キー，アテンション重みと呼ばれ，A の $(i,\ j)$ 成分が大きい場合，文中の i 番目のトークンが j 番目のトークンに注意を向けると解釈される．

3 位置に基づく注意のメカニズム

本節では，まず，各ヘッドは注意の方向と強さに基づいて分類できることを k-means 法を用いて確認する（3.1 項）．次に，位置埋め込みは学習によって周期性を獲得しており（3.2 項），自己注意機構内のクエリとキーにも同じ周波数の成分が含まれることを示す（3.3 項）．最後に，クエリとキーを比較すると，振幅スペクトルは類似しているが位相にズレがあることについて分析し（3.4 項），このズレの方向が注意の向きを決定づけることを示す（3.5 項）．

分析には，事前学習済みモデルである roberta-base を用いた．[1)] モデルの構造は 12 層で各層にヘッドは $n = 12$ 個あり，$T = 512$, $d = 768$ である．また，図を作成する際は文献 [1] の 1 章までの本文を用いた．

3.1 自己注意機構の位置依存性の概要

アテンション重み A_{lh} を可視化すると，例えば，$A_{8,9}$, $A_{2,3}$ では各トークンは隣のトークンに強く注意を向け，$A_{1,1}$, $A_{7,5}$ では左右の周辺トークンへ注意を緩やかに偏らせていることが分かる（図 2）．

1) https://huggingface.co/roberta-base

図 1 ヘッドごとに，100 個の a_{lh} に振られたラベルの割合を円グラフで図示した．

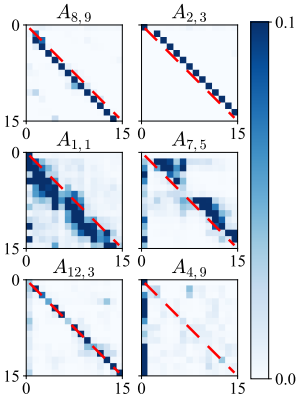

図 2 15 トークン目までのアテンション重み．赤点線は対角線を示す．

ヘッドの注意の傾向を調べるため，各アテンション重み A_{lh} について，対角成分の t 個右の成分の和

$$\mathrm{tr}_t(A) = \begin{cases} \sum_{i=1}^{512-t} a_{i,i+t} & (t \geq 0) \\ \sum_{i=1}^{512+t} a_{i-t,i} & (t < 0) \end{cases} \tag{7}$$

を $t = -10, \ldots, 10$ について並べたベクトル $a_{lh} = [\mathrm{tr}_{-10}(A_{lh}), \ldots, \mathrm{tr}_{10}(A_{lh})]$ を計算した．ベクトル a_{lh} は，A_{lh} における相対位置による注意の偏りの平均的な傾向を表す．入力として wikitext-2 [7] から 512 トークンの文章を 100 件作成し，$100 \times 12 \times 12$ 個のベクトル a_{lh} に k-means 法を適用した．クラスタ数を 6 とすると，ベクトル a_{lh} は注意の強さと向きに基づいて分類され，ほとんどのヘッドで注意の向きは入力文に依存しないことが確認できた（図 1）．

3.2 位置埋め込みの振幅スペクトル

RoBERTa の位置埋め込み $PE \in \mathbb{R}^{512\times 768}$ にも Transformer と同様にトークン位置方向に周期性があるのかを調べるために，PE の各列ベクトルに対して離散フーリエ変換を適用した:

$$\mathrm{spec}_i = \mathrm{FT}(PE_{1,i},\ PE_{2,i}, \ldots, PE_{512,i}) \quad (i = 1,\ 2, \ldots, 768) \tag{8}$$

得られた PE の各列のスペクトルの振幅を周波数ごとに平均したものを図 3 に示す．振幅スペクトルがいくつかの周波数でピークをとることから，RoBERTa の位置埋め込みは Transformer のように明示的に正弦波を用いて定義されていないにもかかわらず，学習により周期性を獲得していると言える．また，位置埋め込みを加える前の単語埋め込み列の振幅スペクトルにはピークが存在しないため，周期性は位置埋め込み特有の性質である．

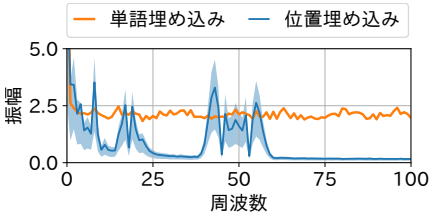

図 3 RoBERTa の位置埋め込みの振幅スペクトル．青線は spec_i の平均で網掛け部分は四分位範囲．オレンジの線は，ある入力に対する，位置埋め込みを加える前の単語埋め込みのスペクトル．

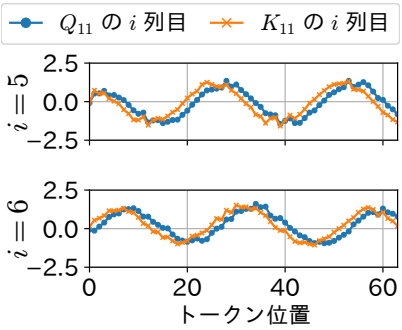

図 4 クエリ $Q_{1,1}$ と キー $K_{1,1}$ の 5, 6 列目の先頭 64 行目までの値．

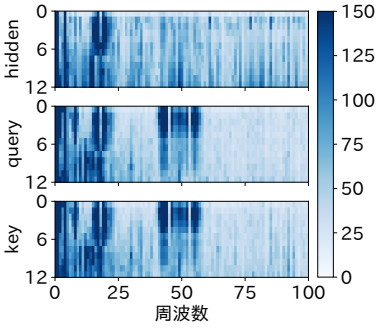

図 5 12 個のヘッドの X, Q, K の振幅スペクトルを，各周波数について最大値をとって層ごとに図示した．

3.3 クエリとキーの振幅スペクトル

特異値分解を用いてクエリとキーを再定義することで，自己注意機構が位置埋め込みに類似した周期的な成分を隠れ状態から抽出できることを示す．

ヘッドのパラメータ W^Q, W^K は $W^A = W^Q(W^K)^T$ とおけば，$XW^Q(XW^K)^T = XW^AX^T$ とまとめられるため，積 W^A のみが意味を持つ．つまり，W^Q, W^K ないし，これらと X の積であるクエリとキーの役割は相対的なものである．実際，クエリとキーの関係を分析する際はパラメータを以下のように再定義すると結果が解釈しやすいことが分かった．すなわち $W^A_{lh} = U^Q_{lh}S(U^K_{lh})^T$ (S は対角行列) と特異値分解して得られる左右特異ベクトルを並べた行列 U^Q_{lh}, U^K_{lh} を用いて，$\mathrm{head}(l,h)$ のクエリとキーをそれぞれ $Q_{lh} = X_lU^Q_{lh}$, $K_{lh} = X_lU^K_{lh}$ と再定義する．ここで，$XW^AX^T = XU^QS(U^K)^TX^T = QSK^T$ である．この Q, K の各列の成分を可視化すると，注意が相対位置に依存するヘッドでは約 4〜12 個の列でトークン位置に関して周期性が見られた（図 4，付録 B）．

各層について，隠れ状態 X_l，クエリ Q_{lh}，およびキー K_{lh} の振幅スペクトルを 12 ヘッド・64 次元に渡って計算して，各周波数における振幅の最大値を求めた（図 5）．すると，位置埋め込みで見られた周波数 50 付近で振幅スペクトルが大きくなる現象が，隠れ状態では見られないがクエリとキーでは見られた．つまり，自己注意機構は位置埋め込み由来の周期的な成分を隠れ状態から抽出できると考えられる．しかし，層を通過するにつれて周波数 50 付近の振幅スペクトルは減衰することから，位置埋め込みの影響は次第に弱まり，推論が位置に依存しにくくなると考えられる．この現象は，下層では構文的特徴を，上層では意味的特徴を学習するという先行研究に対応していると予想される [8].

3.4 位相のズレと周波数の関係

図 4 を見ると，クエリとキーの位相がずれていることが分かる．Vaswani らの推測通り，この現象は回転行列によりクエリとキーがトークン位置方向に相対的にずらされた結果であることを示す．

まず，U^K をある行列 R を用いて $U^K = U^QR$ と分解する．つまり，隠れ状態 X に U^Q を掛けてクエリ $Q = XU^Q$ を作成した後に，さらに R で変換したものがキー $K = XU^K = QR$ であると解釈し，R をもとにクエリとキーの関係性を調査する．

特異ベクトルの直交性より，R の成分は左右の特異ベクトルの内積であり，R も直交行列となる．

$$R = (U^Q)^{-1}U^QR = (U^Q)^{-1}U^K = (U^Q)^TU^K \quad (9)$$

上式で求めた R を対角化して固有値 λ_i を得る．

$$\Lambda = P^{-1}RP = P^{-1}(U^Q)^{-1}U^KP = (U^QP)^{-1}(U^KP)$$
$$\text{ただし}, \Lambda = \mathrm{diag}(\lambda_1, \cdots, \lambda_{768}) \quad (10)$$

直交行列の固有値は $\lambda_i = \cos\theta_i \pm j\sin\theta_i$ (j: 虚数単位) という形で表せるので，固有値の偏角を求めることで R が 768 次元空間をどのように回転させる行列なのかが分かる．式 (3) より，回転行列を掛けることは正弦波の位相をずらす変換であるため，偏角 θ_i は，固有値 λ_i に対応する固有ベクトル $\boldsymbol{p}_i$ によってクエリおよびキーから抽出された 2 つの波形 $Q\boldsymbol{p}_i$, $K\boldsymbol{p}_i \in \mathbb{C}^{512}$ の間の位相のズレである．

偏角 θ_i の単位はラジアンなので，トークン数を単位とする位相のズレ Δ を知るためには，各波形の周波数 f を用いて $\Delta(f, \theta) = 512\theta_i/2\pi f$ を計算する必要がある．そこで，$Q\boldsymbol{p}_i$ の振幅スペクトルを $FT(Q\boldsymbol{p}_i) \in \mathbb{R}^{512}$ としたときの，2 変数関数 $g(f, \theta_i) = FT(Q\boldsymbol{p}_i)_f$ を（$K\boldsymbol{p}_i$ についても同様に）可視化することで周波数と位相の関係を分析した．

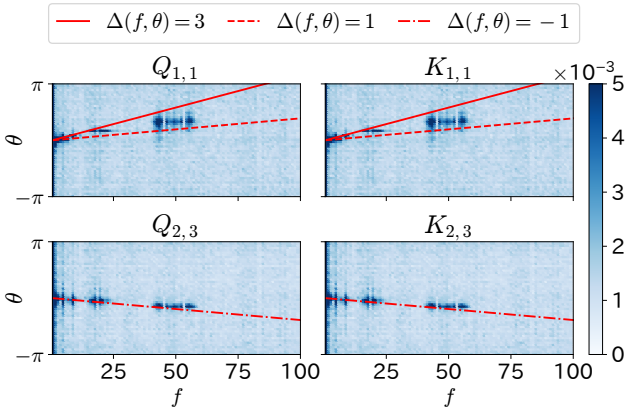

図 6　2 変数関数 $g(f,\theta)$ の 2 次元ヒストグラム．

すると，$g(f,\theta)$ のピークは，左に弱く注意が偏る head(1, 1) では $1 \le \Delta \le 3$ の領域内に，右隣に強く注意が集中する head(2, 3) では $\Delta = -1$ の直線上に分布することが分かった（図 6）．

以上より，クエリとキーの違いを生じさせる行列 R は 768 次元空間内の複数の軸方向に関して異なる回転角を学習するが，周波数成分に分解してみるとトークン単位では一定の個数だけ位相がずれるように学習され，ずれるトークン数と向きはアテンション重みにおいて注意が向く位置に対応することが分かった（図 2）．自己注意機構のパラメータに回転行列を導入した RoPE [9] を用いると学習の収束が早まるのは，本項で述べた性質をより早く獲得するからであると考えられる．

3.5　位相のズレとアテンション重みの関係

2 信号間の時間遅れの計測等に用いられる相互相関と相互共分散を用いて，トークン数単位でクエリとキーの位相のズレを定量化し，クエリとキーの位相のズレの方向に注意が平均的に偏ることを示す．

位相のズレを定量化するために，クエリとキーの各列の相互相関 xcorr と相互共分散 xcov を計算した．ただし，クエリとキーに存在する周期性のみに焦点を当てるため，相互相関に中心化を適用した（式 12）．また，Q と K の j 列目を $\boldsymbol{q}_j$ および $\boldsymbol{k}_j$，Q と K の (i,j) 成分を $q_{i,j}$ および $k_{i,j}$ とする．

$$\mathrm{xcov}_j(t) = \begin{cases} \sum_{i=1}^{512-t} q_{i,j} k_{i+t,j} & (t \ge 0) \\ \sum_{i=1}^{512+t} q_{i-t,j} k_{i,j} & (t < 0) \end{cases} \tag{11}$$

$$\mathrm{xcorr}_j(t) = \frac{\mathrm{xcov}_j(t) - \mathbb{E}_t[\mathrm{xcov}_j(t)]}{\|\boldsymbol{q}_j\| \|\boldsymbol{k}_j\|} \tag{12}$$

相互共分散を各列 j に対応する特異値 s_j で重み付けた和は，QSK^T の対角成分の t 個右の成分の和

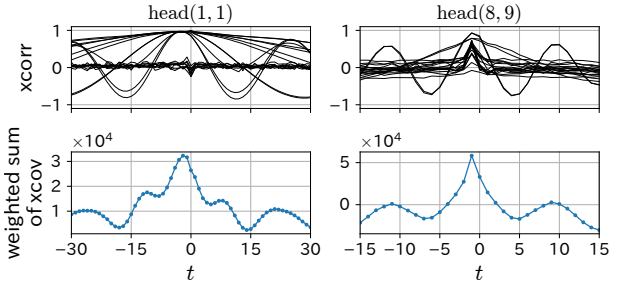

図 7　上: $\mathrm{xcorr}_j(t)$ $(j = 1, \ldots 20)$．下: $\sum_j^{64} s_j \mathrm{xcov}_j(t)$

$\mathrm{tr}_t(QSK^T)$ と等しいことから，注意は位相のズレの方向に平均的に集中する．

$$\mathrm{tr}_t(QSK^T) = \sum_{i=1}^{512-t} (QSK^T)_{i,i+t} \quad (\because \text{式 (7)}) \tag{13}$$

$$= \sum_{i=1}^{512-t} \sum_{j=1}^{64} s_j q_{i,j} k_{i+t,j} \tag{14}$$

$$= \sum_{j=1}^{64} s_j \mathrm{xcov}_j(t) \tag{15}$$

注意の集中が相対位置に基づくヘッドでは，$\boldsymbol{q}_j$，$\boldsymbol{k}_j$ の相互相関は複数の列 j で周期的になり，相互共分散の重み付き和は，アテンション重みと同様に head(1, 1) では $t = -2$ 付近で緩やかなピークをとり，head(8, 9) では $t = -1$ で鋭いピークをとる（図 2, 7）．

図 7（左）において，相互共分散の重み付き和を求める際，$t = -2$ では極大値が重なるためピークは増幅されるが，$t = \pm 30$ 付近では極大値と極小値が重なるためピークは減衰する．その結果，相互共分散の重み付き和は $t = -2$ で最大となったと考えられる．一方，$t = -1$ でのみピークをとる孤立波のような成分も存在し，正弦波状の相互共分散の和が形成する緩やかなピークを強調するような効果が確認された（図 7 右）．この成分を，位置埋め込み由来の周期的な成分へとさらに分解できるかは不明である．

以上より，注意が自身の近傍のトークンに集中する現象は，クエリとキーをトークン位置に関する波形と見たときの位相のズレによって引き起こされていると考えられる．

4　おわりに

自己注意機構において注意の向きが相対位置に依存する現象は，クエリとキーに存在する位置埋め込み由来の周期的な成分の位相が相対的にずらされたためであることを示した．これより，絶対位置埋め込みで相対位置に基づいた推論ができるという経験的な事実のメカニズムの一部が明らかになった．

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. Curran Associates, Inc., 2017. https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

[2] Jesse Vig. A multiscale visualization of attention in the transformer model. In **Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations**, pp. 37–42. Association for Computational Linguistics, July 2019. https://www.aclweb.org/anthology/P19-3007.

[3] Benjamin Hoover, Hendrik Strobelt, and Sebastian Gehrmann. exBERT: A Visual Analysis Tool to Explore Learned Representations in Transformer Models. In **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations**, pp. 187–196, Online, July 2020. Association for Computational Linguistics. https://aclanthology.org/2020.acl-demos.22.

[4] Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. What does BERT look at? an analysis of BERT's attention. In **Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP**, pp. 276–286. Association for Computational Linguistics, August 2019. https://aclanthology.org/W19-4828.

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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)**, pp. 4171–4186. Association for Computational Linguistics, June 2019. https://aclanthology.org/N19-1423.

[6] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach, 2019. https://arxiv.org/abs/1907.11692.

[7] Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. In **International Conference on Learning Representations**, 2017. https://openreview.net/forum?id=Byj72udxe.

[8] Ian Tenney, Dipanjan Das, and Ellie Pavlick. BERT rediscovers the classical NLP pipeline. In **Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics**, pp. 4593–4601, Florence, Italy, July 2019. Association for Computational Linguistics. https://aclanthology.org/P19-1452.

[9] Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding, 2021. https://arxiv.org/abs/2104.09864.

[10] Yu-An Wang and Yun-Nung Chen. What do position embeddings learn? an empirical study of pre-trained language model positional encoding. In **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)**, pp. 6840–6849, Online, November 2020. Association for Computational Linguistics. https://aclanthology.org/2020.emnlp-main.555.

A　様々なモデルの位置埋め込み

Hugging Face に公開されている事前学習済みモデルから絶対位置埋め込みをもつモデルをいくつか選び，図 3 (roberta-base) と同様に位置埋め込みの振幅スペクトルを図示した．

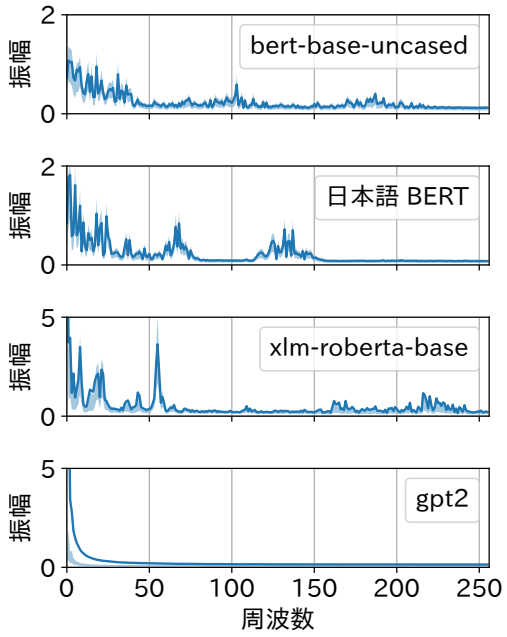

図 8　位置埋め込みの振幅スペクトル．日本語 BERT は cl-tohoku/bert-base-japanese-whole-word-masking である．

エンコーダモデルである BERT と RoBERTa では振幅スペクトルにピークが見られ，RoBERTa の方がピークにおける振幅が大きい．一方，デコーダモデルである GPT-2 は低周波成分しか存在しない．これは，事前学習時の目的関数が強く影響しているためであると考えられる [10]．

B　位置埋め込みの主成分について

相対位置に依存して注意が集中するヘッドにおいて，クエリとキーには約 4〜12 個の次元で位置埋め込みに由来すると考えられる周期的な成分がそれぞれ存在した．位置埋め込みに対して主成分分析を行い，主成分の累積寄与率を求めると，累積寄与率は 4 次元までで 50.51%，12 次元までで 92.23% だった．これより，768 次元空間において 4〜12 次元は少なく感じるが，位置埋め込み空間を十分再現していると考えられる（図 9）．

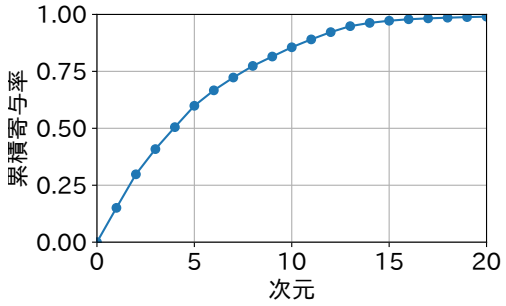

図 9　roberta-base の位置埋め込みの累積主成分寄与率．