部分空間の擬似直交性による Transformer 言語モデルの内部表現の解釈

前田晃弘 [1,4]　鳥居拓馬 [2]　日高 昇平 [1]　井之上直也 [1]　大関洋平 [3]
[1] 北陸先端科学技術大学院大学　[2] 東京電機大学　[3] 東京大学　[4] 日本学術振興会特別研究員
akihiro.maeda@jaist.ac.jp

概要

本研究は，Transformer 言語モデルにおける内部表現を解釈するために，擬似直交性の概念を新たに導入してそのアテンション層および FFN 出力層の部分空間の幾何的関係を分析する．FFN 層のウェイト行列の行空間が語彙空間と擬似直交していることを示した上で，FFN 層からの出力が意味的概念を担うコンセプトベクトルとして機能し，内部表現の文脈化に寄与している可能性を明らかにする．

1 はじめに

1.1 Mechanistic interpretability

Mechanistic interpretability（MI: 内部機序解釈可能性）[1] と呼ばれる一連の研究では，深層学習モデルの内部表現を解釈するため学習済みニューラルネットワークをリバースエンジニアリングする．パラメータや内部表現を直接数理的に分析して，解釈可能な計算回路や特徴量を特定し，その機序の説明 (mechanistic explanation) を与える．

MI 研究は大規模言語モデル (LLM) において特に活発である [2]．LLM の内部機序には依然として多くの未解明な側面が存在する．その解明は性能改善に加え，自然言語の構造や意味合成 [3] に関する新たな知見をもたらすことが期待される．

1.2 Logit Lens: 単語分布への写像

MI 研究で注目される手法の一つとして，Logit lens [4] がある．これは，LLM の内部表現を単語埋め込み行列を用いて単語分布へ写像する（“logit” を計算する）ことで，人が理解できる表現（単語）に対応づけ解釈する手法である．例えば，[5] は GPT2 の Feed Forward Network (FFN) 層の出力ウェイトへ Logit lens を適用すると意味的な概念を共有する単語群が現れることを実証的に示した．

一方で，Logit lens が有効でない場合も報告され，内部表現を解釈できるよう復号化するための改善方法が研究されている [6]．また，[7] は，FFN 層とは異なる変換を用いて Attention 層の作用を解釈している．これらの観察事実は Transformer の各層が異なる性質の計算を行うことを示唆する．

1.3 部分空間の疑似直交性

単語分散表現に関する先行研究 [8, 9] は，単語ベクトルの集合が代数系としてのベクトル空間をなしており，語義や特徴が線形部分空間に対応する構造にあることを示唆する（線形表現仮説 [10]）．

本研究では，Transformer の各層が異なる線形部分空間を構成している可能性に着眼し，その幾何的な関係を調べる．本研究の新規な試みとして，ノイズを許容して直交性の定義を緩めた擬似直交という概念を適用し，次元数以上の擬似的な直交基底が Transformer 各層の部分空間をなすことを示す．FFN 層が概念を表現するベクトル（コンセプトベクトル）を事前学習しており，その出力が単語表現を文脈化して意味を再構成している可能性を指摘する．

2 Transformer 各層のなす部分空間

2.1 Transformer の概要 [11]

式（1–6）に概要を示す．Transformer は，トークン長 n の入力文を埋め込み行列 $E \in \mathbb{R}^{V\times d}$ により n 個の d 次元ベクトルへ変換した上で（V は語彙サイズ），各トークン位置を符号化したベクトルを加算し，縦結合して当初の内部状態 $X^{l=0}$ とする（式 1）．

$$X^0 = X_e + X_p \in \mathbb{R}^{n\times d} \tag{1}$$

$$Y^l = A^l X^l W_v^l W_o^l =: A^l X^l W_{vo}^l \in \mathbb{R}^{n\times d} \tag{2}$$

$$M^l = \mathrm{LayerNorm}\left(Y^l + X^l\right) \in \mathbb{R}^{n\times d} \tag{3}$$

$$N^l = \mathrm{ReLU}(M^l W_{in}) \in \mathbb{R}_+^{n \times 4d} \tag{4}$$

$$Z^l = N^l W_{out} \in \mathbb{R}^{n \times d} \tag{5}$$

$$X^{l+1} = \mathrm{LayerNorm}\left(Z^l + M^l\right) \in \mathbb{R}^{n \times d} \tag{6}$$

Transformer の各層（$l = 0, \ldots, L-1$）は，アテンション層（式 2）と FFN 層（式 4,5）と呼ばれる計算ユニットを交互に適用する．$W_q^l, W_k^l, W_v^l, W_o^l, W_{in}^l, W_{out}^l$ は各層で学習されるパラメータであり，A^l はアテンションウェイトの結合を表す[1)]．式（3,6）の LayerNorm はレイヤー正規化 [12] を行う．$\mathrm{ReLU}(x) = max(x, 0)$．最終層後の内部状態 X^L は再変換のための行列（多くの場合，埋め込み行列の転置行列 $E^\top$）により単語へ変換される．レイヤー正規化を無視すると，

$$X^L = Z^{L-1} + M^{L-1} = Z^{L-1} + (Y^{L-1} + X^{L-1}) \tag{7}$$

$$= (Z^{L-1} + Y^{L-1}) + Z^{L-2} + M^{L-2} = \cdots \tag{8}$$

$$= \sum_{l=0}^{L-1} (Y^l + Z^l) + X^0 \tag{9}$$

と式変形でき，当初の内部状態 X^0 にアテンション層の出力 Y^l（Y ベクトルと呼ぶ）と FFN 層の出力 Z^l（Z ベクトルと呼ぶ）を足し込む計算フロー（Residual stream と呼ぶ [6]）と捉えられる．

2.2 各計算ユニットがなす部分空間

式（1–6）で示した計算フローを，BERT を例として，各層の次元数とともに図 1 に示す．内部状態（行列）の各行ベクトルから同じ次元数の Y ベクトルと Z ベクトルへの変換は局所的な空間ごとに見ればアフィン自己準同型写像のように扱える．

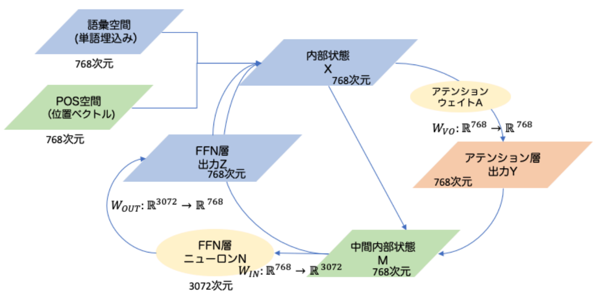

図 1 BERT の計算フロー

Y ベクトルは，式（2）よりアテンションウェイトで加重和された内部状態 $A^l X^l$ を行列 W_{VO}^l により線形写像した像と見れる（行ベクトルに右から行列をかける；バイアスを無視）．この線型写像の像は W_{VO}^l の行空間であり，任意の内部状態を変換して得られる Y ベクトルは全てこの部分空間にある．

同様に，Z ベクトルは式（5）より行列 W_{out}^l により表現される線型写像の像である部分空間の中にある．すなわち，Z ベクトルは W_{out}^l の行ベクトルを N^l の行ベクトルの成分により線形結合したものであり，W_{out}^l の行空間に住む．N^l は活性化関数の出力であり（式 4），W_{out}^l の行ベクトルの活性・非活性を制御するのでニューロンと呼ばれる [6]．BERT では各層のニューロンは $4d = 3072$ 個である．ニューロンはスパースとなるが次節で述べる擬似直交と密接な関係がある [13]．

X^0（式 1）は，単語埋め込み行列 E と位置行列 P の行ベクトルの和であり，それぞれのベクトルが張る部分空間を語彙空間と POS 空間と呼ぶ．

2.3 擬似直交性

前節で見た各部分空間の間の幾何的関係を特徴づけるために擬似直交性を導入する（定義は次節）．二つのベクトルが直交するとは，その内積（またはコサイン類似度）がゼロであることだが，ノイズを許容しコサイン類似度が $0 \pm \epsilon$ の範囲にあることを擬似直交と呼ぶこととする．d 次元のベクトル空間にある任意のベクトルは d 個の（直交）基底の線型結合に分解できるが，ノイズを許容すれば d よりもはるかに大きい個数のベクトルを基底のように用いた擬似的な直交分解が可能となる [14]（付録 A).

[13] は擬似直交な基底が解釈可能な特徴や意味を表現しており，深層学習モデルの内部表現は擬似的な直交基底の重ね合せであるとの立場を示している（superposition 仮説）．[13] はトイモデルを用いて，学習データがスパースな場合，モデルが擬似直交を用いた表現を学習して，内部表現の次元数以上の特徴量を表現できることを示している．

3 分析：Transformer の部分空間

3.1 擬似直交性を用いた分析手法

分析の目的 Transformer の各層ウェイト行列の行空間の直交性の程度（擬似直交性）を評価する．

定義 d 次元のベクトル空間 V の 2 つの部分空間 $U, W \subseteq V$ が擬似直交するとは，正規化された任意のベクトル $\mathbf{u} \in U, \mathbf{w} \in W$ に対して，$-\epsilon \leq \langle \mathbf{u}, \mathbf{w} \rangle \leq \epsilon$ で

1) アテンションヘッドと呼ばれるサブユニットで以下の計算を行う．W_q^l, W_k^l を分割した $W_q^{l,h}, W_k^{l,h}$ を用いて，まずヘッドごとのウェイト $A^{l,h} = \mathrm{Softmax}\left(X W_q^{l,h} W_k^{l,hT} X^T\right) \in \mathbb{R}^{n \times n}$ を計算する．ソフトマックスは行方向に総和が 1 となるよう適用，その上で次の結合を行う．$A^l X^l W_v^l W_o^l := [A^{l,1}, \cdots, A^{l,12}] X^l [W_v^{l,1}, \cdots, W_v^{l,12}]^T W_o$

あることとする．これは次式と同値である．

$$\mu_w := \max_{\mathbf{u}\in U} |\cos(\mathbf{u}, \mathbf{w})| \leq \epsilon \quad (\forall \mathbf{w} \in W) \tag{10}$$

μ_w は $\mathbf{w} \in W$ に対して計算され，$\mathbf{u} \in U$ を動かした時の最大絶対コサイン類似度である．「完全に」擬似直交する場合には，すべての $\mathbf{w} \in W$ に対して μ_w が $\pm\epsilon$ の範囲内である．μ_w が $\pm\epsilon$ の範囲内にある $\mathbf{w}$ の比率を「擬似直交の度合い」とする．

分析手順 与えられた二つのベクトル群の一方を式（10）における U，他方を W として，W に含まれる全ての行ベクトル $\mathbf{w}$ に対して μ_w を算出する．アテンション層と FFN 層の解釈可能性を調べるために，それぞれの部分空間が語彙空間とどの程度擬似直交しているか評価する必要があるので，それらを W, 語彙空間を U として式（10）を適用する．

データ Huggingface の事前学習済みモデル（Pytorch 版）[15] より BERT-base と GPT2 を用いる．いずれも 12 層の Transformer モデルであり，前者は Bidirectional encoder，後者は Unidirectional decoder である．それぞれ単語埋め込み行列 E, 位置ベクトル行列 P 並びにアテンション層の行空間を与える行列として W^l_{vo} と FFN 層の行列として W^l_{out} を用いる．

3.2 予備分析：位置ベクトルと語彙空間

BERT が用いる単語埋め込み行列 E は，30522 個の 768 次元ベクトルからなる．すべての組のコサイン類似度の分布は図 2 左の通りほとんどが正で中央値は 0.447 である（クラスタリングすると，同一クラスタの単語は意味的なまとまりを持つ [16]）．これらのベクトルの張る部分空間が語彙空間である．

BERT の 512 個ある位置ベクトルは語彙空間とどのような幾何的関係にあるだろうか．単語埋め込みと位置ベクトルのコサイン類似度（30522×512 組）は，ほぼすべて ± 0.1 の範囲内にある．（図 2 右）

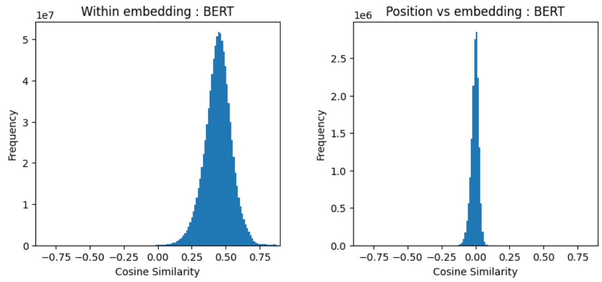

図 2 単語埋め込みと位置ベクトルのコサイン類似度

POS 空間と語彙空間の擬似直交の度合いを評価するために，定義式（10）により位置ベクトルを W として算出した μ_w の値を図 3 右に昇順に示す．横軸は個々の位置ベクトルの順位に対応する．μ_w の最大値は図の最右で 1.0 であるが，二番目は 0.174 であり，$\epsilon = 0.15$ とする時 $\mu_w \leq \epsilon$ となる w の比率 95.9%が POS 空間と語彙空間の擬似直交の度合いとなる．$\mu_w = 1.0$ となる組み合わせはトークン [CLS] と位置 0 である[2]．POS 空間は [CLS] を除く語彙空間と擬似直交している．GPT2 に関する同様の分析

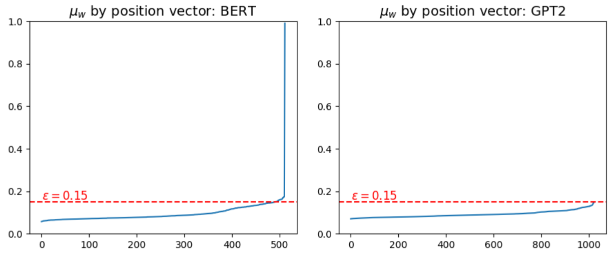

図 3 POS 空間と語意空間の擬似直交性

（図 3 右）は，すべての位置ベクトルが語彙空間と擬似直交することを示す．1052 個の位置ベクトルに対して算出された μ_w の最大値は 0.148 である．

3.3 分析結果：アテンション層と FFN 層

アテンション層と FFN 層について，その部分空間と語彙空間との擬似直交性を調べる．式（10）に従い，単語埋め込み行列を U として，アテンション層に関しては行列 $W^l_{VO}(l = 0, \ldots, 11)$ の行ベクトル，FFN 層に関しては $W^l_{out}(l = 0, \ldots, 11)$ の行ベクトルに対して，値 μ_w を算出した．図 4 は BERT の各 l

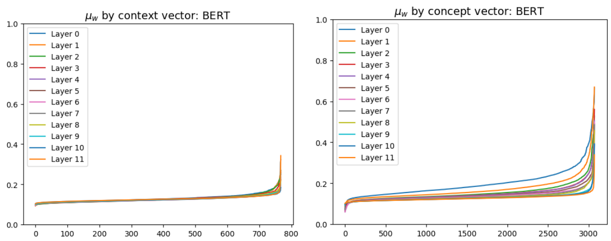

図 4 アテンション層（左）と FFN 層（右）の語彙空間との擬似直交の度合い（層 $l = 0, \ldots, 11$ ごと）：BERT-base

層ごとにアテンション層（左）と FFN 層（右）について値 μ_w を昇順に示している．$\epsilon = 0.15$ として，アテンション層では 12 層平均で 93.2%が語彙空間と擬似直交する一方，FFN 層では平均で 71.8%の W_{out} の行ベクトルが語彙空間と擬似直交である．逆に言えば，29.2%において単語埋め込みとのコサイン類似度が ϵ 以上ということである．但し，層によるば

2） BERT では文頭記号としてトークン [CLS] が追加され，その文中の位置は常に 0 となることから，[POS0] が [CLS] と一致されるよう学習されていることが示唆される．

らつきが大きく低層ほど語彙空間と共有する部分空間が大きい（$l = 0$ では 80.0%，$l = 5$ では 26.4%）．

同様の傾向が GPT2 でも観察された（図 5）．アテンション層（左）ではほとんどの層において語彙空間と擬似直交する一方，FFN 層（右）は高層において語彙空間との交わりをもつ．$\epsilon = 0.2$ とした時，アテンション層の行ベクトルのうち 88.8%（11 層を除くと 92.7%）が擬似直交であるに対して，FFN 層では全層平均では 86.1%が擬似直交であるが，6 層以上に限ると平均 77.1%，10 層以上の平均は 67.5%と高層ほど語彙空間に近い．BERT との違いは，FFN 層において低層と高層で語彙空間との擬似直交の度合いが逆転することと，アテンション層の最終層のみ語彙空間とは擬似直交でないことである．エンコーダとデコーダの違いと考えられるが，その機序の解明は今後の研究課題である．

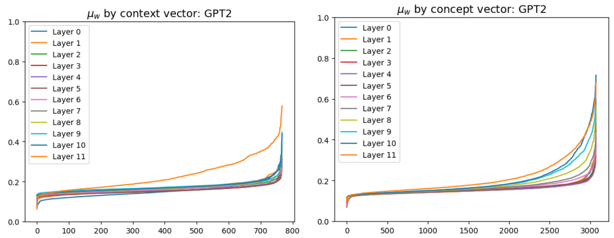

図 5 語彙空間との擬似直交の度合い：GPT2

4 コンセプトベクトルによる文脈化

4.1 Logit lens が有効な部分空間

擬似直交分析からの示唆は，アテンション層の行空間が語彙空間とほぼ直交する一方，FFN 層はそうではなく，一定数の行ベクトルが単語埋め込みベクトルのいずれかと非ゼロの内積をとることである．実際に Geva ら [5] は FFN 層から抽出した行ベクトル（本研究ではコンセプトベクトルと呼ぶ）を単語分布に変換し，その 20–40%が解釈可能な意味のまとまり持つことを報告している．本研究の分析は，Transformer において Logit lens が有効な部分空間（FFN 層）とそうでない部分空間（アテンション層）の存在に説明を与えるものであり，Geva らの研究に数理的根拠を与える．

4.2 内部表現の解釈と文脈化の過程

Z ベクトルがコンセプトベクトルの，Y ベクトルがアテンション層の行ベクトルのそれぞれ線形結合であること，またアテンション層と POS 空間が語彙空間と擬似直交していると見做せることから，式（9）に Logit lens を適用して次式を得る．

$$X^L E^T = \left(X_e + X_p + \sum_{l=0}^{L-1} Y^l + \sum_{l=0}^{L-1} Z^l \right) E^T \approx \left(X_e + \sum_{l=0}^{L-1} Z^l \right) E^T \tag{11}$$

同式は，入力文の単語埋め込みに対して，各層の Z ベクトルを順に加算することで最終層の出力を得る文脈化のプロセスと解釈することができる．検証のため BERT へ以下の入力文を与えて，FFN 各層出力の Z ベクトルの総和をとり，Logit lens を適用する．

文 1 “The king wears a tie”
文 2 “He ties the package with string before mailing it”
文 3 “The match ended in a tie, so both teams shared the points”

表 1 に文 1 から得られた総和 Z ベクトルに最も近い単語を位置ごとに示す．位置番号 3 を除き，入力文の単語が top に来ており入力文を復号している．

表 1 コンセプトベクトルによる再構成

位置番号	0	1	2	3	4	5	6
入力文	[CLS]	the	king	wears	a	tie	[SEP]
Top	**[CLS]**	**the**	**king**	is	**a**	**tie**	**[SEP]**
類似度	.32	.39	.34	.30	.45	.16	.28

Top 以外の上位単語からは，総和 Z ベクトルが符号化している文脈を観察することができる．表 2 は，3 文に共通する多義語 *tie* に対応する総和 Z ベクトルを変換した際の上位単語を示す．文脈に沿った単語群が上位に現れており，総和 Z ベクトルが文脈依存の語義を反映していることを示唆する．

表 2 総和 Z ベクトルに現れる多義語の文脈

順位	語義 1	語義 2	語義 3
文脈	衣類（名詞）	縛る（動詞）	引分け（名詞）
1	tie	tied	tie
2	ties	closed	tied
3	clothes	tying	game
4	plaid	cut	tying
5	jacket	worked	rivalry
6	shirts	tie	comparison
7	wardrobe	cutting	conflict

5 まとめと考察

語彙空間との擬似直交性の違いから各層に対する Logit lens の有効性の違いを説明し，FFN 層の出力が Transformer の内部状態の文脈化を担っている可能性を示した．今後の課題は，アテンションヘッドによるコンセプトベクトルの活性化分析により意味合成の機序を解明することと，アテンション層が意味合成の文脈を準備するプロセスを条件付き確率としてモデル化することである．

謝辞

本研究は科研費基盤研究 B(一般) JP23H0369，JST さきがけ JPMJPR20C9，JST CREST JPMJCR23P4，JSPS KAKENHI 24KJ1202 の助成を受けて行われた．

参考文献

[1] Lena Kästner and Barnaby Crook. Explaining ai through mechanistic interpretability. **European Journal for Philosophy of Science**, Vol. 14, No. 4, p. 52, 2024.

[2] Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. Progress measures for grokking via mechanistic interpretability, 2023.

[3] Marco Baroni, Raffaella Bernardi, Roberto Zamparelli, et al. Frege in space: A program for compositional distributional semantics. **Linguistic Issues in language technology**, Vol. 9, pp. 241–346, 2014.

[4] nostalgebraist. interpreting gpt: the logit lens, 2020. www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/.

[5] Mor Geva, Avi Caciularu, Kevin Ro Wang, and Yoav Goldberg. Transformer feed-forward layers build predictions by promoting concepts in the vocabulary space. **arXiv preprint arXiv:2203.14680**, 2022.

[6] Javier Ferrando, Gabriele Sarti, Arianna Bisazza, and Marta R. Costa-jussà. A primer on the inner workings of transformer-based language models, 2024.

[7] Guy Dar, Mor Geva, Ankit Gupta, and Jonathan Berant. Analyzing transformers in embedding space. **arXiv preprint arXiv:2209.02535**, 2022.

[8] Tomas Mikolov. Efficient estimation of word representations in vector space. **arXiv preprint arXiv:1301.3781**, Vol. 3781, , 2013.

[9] Sanjeev Arora, Yuanzhi Li, Yingyu Liang, Tengyu Ma, and Andrej Risteski. Linear algebraic structure of word senses, with applications to polysemy. **Transactions of the Association for Computational Linguistics**, Vol. 6, pp. 483–495, 2018.

[10] Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models, 2024.

[11] 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 I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, **Advances in Neural Information Processing Systems**, Vol. 30. Curran Associates, Inc., 2017.

[12] Jimmy Lei Ba. Layer normalization. **arXiv preprint arXiv:1607.06450**, 2016.

[13] Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, Roger Grosse, Sam McCandlish, Jared Kaplan, Dario Amodei, Martin Wattenberg, and Christopher Olah. Toy models of superposition, 2022.

[14] Pentti Kanerva. Hyperdimensional computing: An introduction to computing in distributed representation with high-dimensional random vectors. **Cognitive computation**, Vol. 1, pp. 139–159, 2009.

[15] huggingface. Pytorch-transformers, 2019. https://huggingface.co/transformers/v1.2.0/index.html.

[16] 前田晃弘, 鳥居拓馬, 日昇平, 大関洋平. 部分空間法に着想を得た transformer のアテンションヘッドにおける特徴抽出. 言語処理学会第 30 回年次大会, 2024.

[17] Benyamin Ghojogh, Ali Ghodsi, Fakhri Karray, and Mark Crowley. Johnson-lindenstrauss lemma, linear and nonlinear random projections, random fourier features, and random kitchen sinks: Tutorial and survey. **arXiv preprint arXiv:2108.04172**, 2021.

[18] Sanjoy Dasgupta and Anupam Gupta. An elementary proof of a theorem of johnson and lindenstrauss. **Random Structures & Algorithms**, Vol. 22, No. 1, pp. 60–65, 2003.

A 擬似直交ベクトルの上限数

d 次元ベクトル空間が含む互いに擬似直交なベクトルの数の試算を試みる．次元数 d を固定したとき，コサイン類似度が閾値 ϵ 以下のベクトル群（擬似直交ベクトル）の最大数 N として，Transformer が利用する擬似直交基底の数を推計する．

A.1 Johnson-Lindenstrauss 補題

先行研究 [13] で引用されるこの補題は，高次元空間にある任意の N 個の点集合を次元削減により低次元空間 $\mathbb{R}^d$ へ埋め込む際に，その点間距離が $1 \pm \epsilon$ の範囲内で近似的に保存されることを保証する．

補題の概略（[17] 定理 1）は，D 次元ベクトルの集合を $\chi = \{x_i \in \mathbb{R}^D\}_{i=1}^N$，エラー許容率を $0 \le \epsilon \le 1$ として，正の整数 d が

$$d \ge \Omega\left(\epsilon^{-2}\log(n)\right) \tag{12}$$

を満たす時，ランダムに生成した写像（表現行列の成分が独立同一な標準正規分布に従う写像）$f : \mathbb{R}^D \to \mathbb{R}^d$ のもとで，ある確率で

$$(1-\epsilon)\|x_i - x_j\|_2^2 \le \|f(x_i) - f(x_j)\|_2^2 \le (1+\epsilon)\|x_i - x_j\|_2^2 \tag{13}$$

が成り立つというものである．Ω はアルゴリズム計算量の漸近的な下限を与える記号である[3)]．この定理の主張は，次元数 d を増やせば，任意のベクトル間の内積を任意の閾値 ϵ に抑えることが可能であることを示唆する．そして，次元数 d が十分に大きければ，ランダムな埋め込みを用いてベクトル間の相対関係（距離と内積）を保存することを意味する．d, N, ϵ の関係式（12）は d を固定した場合，擬似直交ベクトル数 N の上限が ϵ に依存して増加することを示す．

A.2 ランダム抽出による数値実験

Johnson-Lindenstrauss 補題は，ランダム写像により次元削減を行う際の全ての点群の挙動を記述するものであり，特に極値（例外的に大きな，あるいは小さな値）が閾値内に保存されることを保証する．そこで具体的に極値の挙動を調べるために，次の数値実験を行った．まず $d = 768$ 次元のベクトル $\boldsymbol{v} \in \mathbb{R}^d$ をランダムに，ベクトルの各成分は独立同一な標準正規分布に従うよう，N 個生成した上で，ノルム 1 に正規化する．N 個のベクトル間の内積 (正規化しているのでコサイン類似度と同じ）の絶対値の最大 $\max_{i \ne j \in [N]} |\langle \boldsymbol{v}_i, \boldsymbol{v}_j \rangle|$ を求める．$N = 10000, 20000, 30000$ の各条件で 100 回の試行を行い，その統計分布を記録した．

図 6 は，各試行ごとの内積最大値を昇順に表示している．ベクトル数 N が増加するに伴い，内積最大値がグラフが上方にシフトする傾向が確認できる．これは極値を得る分布のサンプル数 $_N C_2$ が増加するに伴い，極端な値が大きくなることを反映している．

本実験は，Johnson-Lindenstrauss 補題のランダム写像の代わりにターゲット空間のベクトルを直接ランダム抽出したものであり，同補題の主張する事実を直接再現するものではない，しかし，両者は高次元空間におけるランダムプロセス用いており，いずれもランダムベクトル間の内積が高次元空間では分布的に集中するという現象を反映している．

3) [定義] ある関数 $g(n)$ について $\Omega(g(n))$ とは，ある正の定数 c, n_0 が存在して，全ての $n \ge n_0$ に対して $0 \le c(g(n)) \le f(n)$ を満たすような $f(n)$ の集合である．

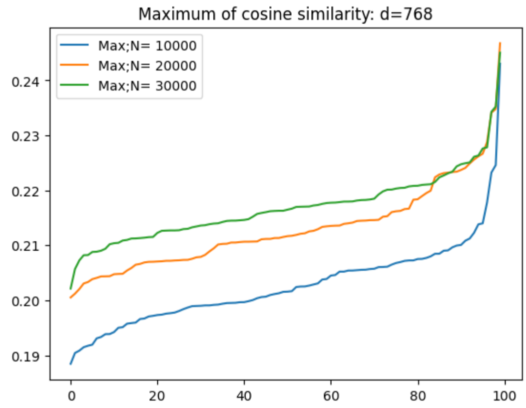

図 6 極値（すべてのベクトル間の内積最大値）の分布

A.3 上限数の理論的推計

Transformer の内部状態に用いるベクトル数は，BERT の場合, 単語埋め込み (30,522 個)，位置ベクトル（512 個），アテンション層の出力ウェイト（768 個 ×12 層）FFN 層の出力ウェイト（3,072 個 ×12 個）であり合計 77,114 個である．GPT2 では，単語埋め込み (50,257 個)，位置ベクトル（1,024 個）でそれ以外は BERT と同じで，合計 97,361 個である．

ここで，$d = 768$ の空間で $N = 100,000$ のベクトルを擬似直交とするための ϵ の理論値を次のように求める．式 (12) の条件を満たす ϵ^* として，[18]（定理 2.1) で提案されている $\Omega(\epsilon^{-2} \ln n) = 8\epsilon^{-2} \ln(n)$ を用いると次の値を得る．これは閾値の下限である．

$$\epsilon^* = \sqrt{\frac{8 \ln N}{d}} = 0.346 \tag{14}$$

なお，$N = 10000$ では $\epsilon^* = 0.310$ であり，数値実験で見たようにランダムサンプリングしたベクトル群は，Johnson-Lindenstrauss 補題を成立させるための条件を充足している．