自己注意機構のアテンション重みが特定の種類のトークンに集中する現象と外れ値次元の関係

丸田 佳　松崎 拓也
東京理科大学　理学研究科　応用数学専攻
1422541@ed.tus.ac.jp　matuzaki@rs.tus.ac.jp

概要

BERT の自己注意機構は一部の層で [CLS]・[SEP] といった特殊トークンや，カンマ・ピリオドに大きなアテンション重みを割り振るという現象が知られている．一方，BERT の各層での出力ベクトルには他の次元と値の絶対値が大きく離れている次元（外れ値次元）が存在することが知られている．

本研究では，外れ値次元と特定の種類のトークンへのアテンションの集中という現象の関係を定量的に分析する．結果として，一部の層では少数の外れ値次元がアテンションの集中を決める支配的な要因になっており，その影響で特定の種類のトークンに割り振られるアテンション重みが大きくなることを数値的に示す．

1　はじめに

自己注意機構をその中核とする Transformer [1] アーキテクチャは，BERT [2] をはじめとするさまざまな言語モデルで用いられている．自己注意機構は，入力文の各トークンに対し，他のトークンとの関連度（アテンション重み）を計算し，関連が強いトークンを重視するように重みづけを行うことで出力ベクトルを計算している．従って，アテンション重みが平均的に大きいトークンは，他のトークンから大きな注目を集めていることになるため，アテンション重みを観察することで，モデル内部の挙動をある程度理解できる場合がある [3, 4, 5]．しかし，アテンション重みが平均的に大きいトークンは必ずしも文において重要なトークンというわけではない．実際に Clark ら [6] は，BERT におけるアテンション重みは，前半層では [CLS]，中間層では [SEP]，後半層ではカンマ・ピリオドといった内容語としての意味を持たないトークンに大きく割り振られることを示した．Kobayashi ら [7] による，特殊トークンや句読点に対応するベクトルはノルムが小さい傾向にあるため，出力ベクトルを作る際のこれらのトークンからの最終的な寄与は小さいことがわかるという観察もあるが，モデル内部の挙動を知る手がかりとしてアテンションを利用するためには，意味を持たないトークンにアテンションが集まる原因についてさらに理解することが必要である．

一方，BERT の各層が出力するベクトルには他の次元と比較して値の絶対値が非常に大きい次元（以降，「外れ値次元」と呼ぶ）が存在することが知られており，外れ値次元は，事前学習コーパスにおけるトークン頻度との相関があること [8] や，下流タスクの知識をエンコードしていること [9]，位置埋め込みや層正規化からの影響 [10, 11] などが明らかになっている．特に，Puccetti ら [8] は，特定の層，特定の外れ値次元に着目した時に，アテンション重みの大きさと外れ値次元の値が相関している場合があることを示し，特定の種類のトークンへのアテンションの集中と外れ値次元の値との間に関係があることを示唆している．

本研究では，外れ値次元において，特定の種類のトークンは他のトークンとの絶対値の差が大きいことを示す（§3）．さらに，BERT の出力ベクトルの各次元とアテンションの関係を定量的に分析することで，少数の外れ値次元がアテンションを決める上で支配的な影響を及ぼす場合があることを示す（§4）．これら 2 つの事実から，BERT が特殊トークンやカンマ・ピリオドへアテンション重みを大きく割り振る現象は，特定の少数次元の強い影響が原因であることがわかる．

2　準備

既存研究で報告されている以下の 2 つの性質を，言語モデルとして BERT-base-uncased[1)], 入力データ

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

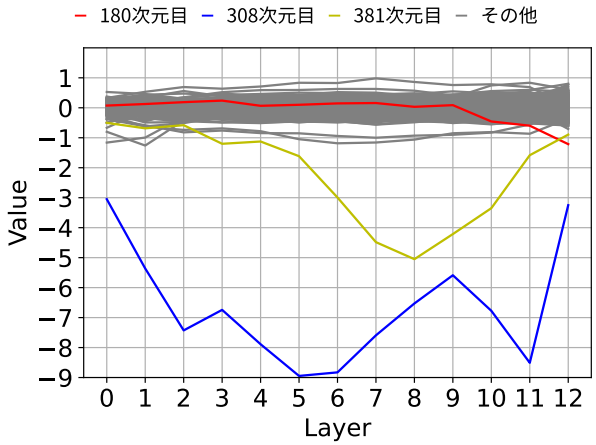


図 1: STS-B dev データに対する BERT の全出力ベクトルの次元ごとの平均値

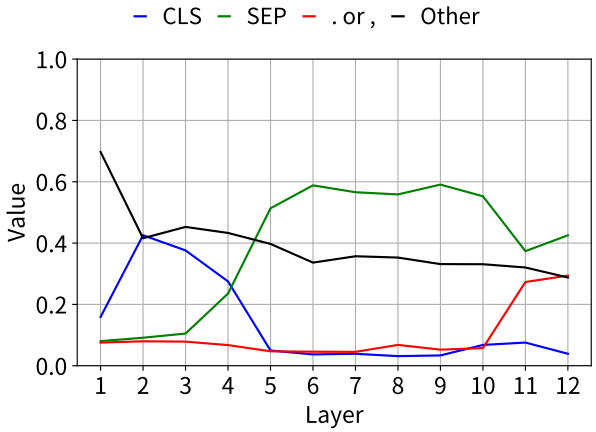


図 2: STS-B dev データに対する BERT のアテンション重みの層平均

として STS-B dev データセット[2] の sentence 1 全文を用いて確認する：

- BERT の出力ベクトルには他の次元よりも絶対値の大きい外れ値次元が存在する [8, 9, 10, 11]
- BERT のアテンション重みは前半・中間・後半層のそれぞれで特定の種類のトークンに大きく割り振られる [6]

2.1 BERT の外れ値次元

全トークンの出力ベクトルを平均し，次元ごとに折れ線グラフにしたものを図 1 に示す．図 1 より，308 次元目は全ての層において絶対値が大きいこと，381 次元目は中間層で他の次元に比べ絶対値が大きくなること，180 次元目は最終層付近で絶対値が大きくなることがわかる．以下，本論文ではこの 3 つの外れ値次元に注目する．

2.2 トークンの種類とアテンション重み

全トークンに対するアテンション重みの平均を [CLS]，[SEP]，カンマ・ピリオド，その他のトークンの 4 つに分けて層ごとに求めたものを図 2 に示す．図より，前半層 (Layer 2，3) では [CLS] へのアテンションが大きく，中間層 (Layer 5〜10) では [SEP] へのアテンションが大きいことがわかる．また，Layer 11，12 においてカンマ・ピリオドへのアテンションが急激に大きくなっていることがわかる．

3 外れ値次元の値とトークンの種類

図 1 で確認した外れ値次元において，特定の種類のトークンは他のトークンとの絶対値の差が大きいことを示す．STS-B の dev データセットの sentence 1 の全文を事前学習済みの BERT-base (uncased) に入力し，各層ごとにトークンを [CLS]，[SEP]，カンマ・ピリオド，その他の 4 種類に分けて，外れ値次元の値の平均値を計算しグラフにまとめたものを図 3a，3b，3c に示す．

308 次元目　図 3a を見ると，Layer 0 から Layer 8 までは [CLS] の絶対値が最も大きく，Layer 0 を除いてカンマ・ピリオドも [CLS] とほぼ同程度の値であることがわかる．また Layer 1 と Layer 2 では [CLS]，[SEP]，カンマ・ピリオドの値は同程度であるが，Layer 3 から [SEP] の値が増加し，Layer 5 以降において [SEP] は非負値をとることがわかる．

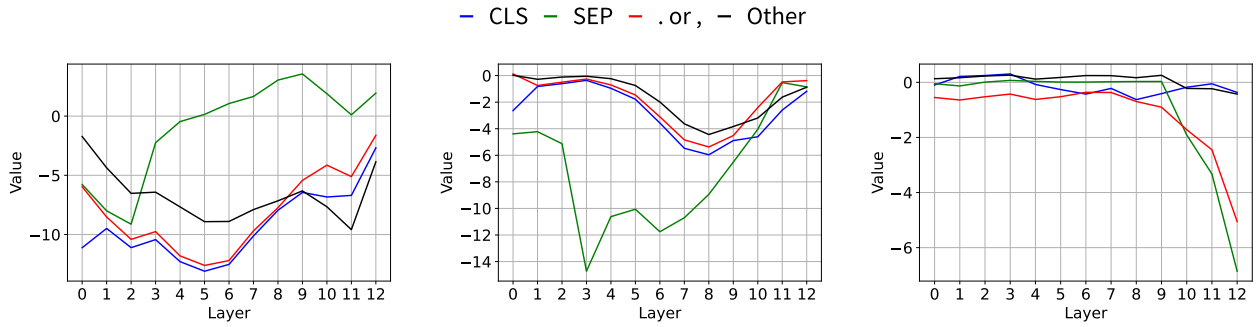
381 次元目　図 3b を見ると，Layer 0 から Layer 8 までは [SEP] の絶対値が大きく，特に中間層においてその傾向が強いことがわかる．

180 次元目　図 3c を見ると，Layer 9 以降から [SEP] とカンマ・ピリオドの絶対値の値が大きくなり始め，最終層では，[SEP] とカンマ・ピリオドは他の種類のトークンと比べて絶対値が非常に大きくなることがわかる．

4 外れ値次元のアテンションスコアへの影響の分析

本節ではアテンションスコアに対する各次元からの寄与の計算方法を述べ，外れ値次元がアテンションスコアにおいて大きな影響を持つことを示す．

2) http://ixa2.si.ehu.eus/stswiki/index.php/STSbenchmark



(a) 308 次元目　(b) 381 次元目　(c) 180 次元目

図 3: 外れ値次元の値の層ごとの平均値

4.1 アテンション重みにおける特定次元からの影響

入力文の m 番目のトークンをクエリ，n 番目のトークンをキーとした時，あるヘッドにおけるトークン m から n へのアテンション重み $\alpha_{m,n}$ は以下のように計算される．

$$\alpha_{m,n} := \mathrm{softmax}\left(\frac{\boldsymbol{q}(\boldsymbol{x}^m)\boldsymbol{k}(\boldsymbol{x}^n)^\top}{\sqrt{d'}}\right) \tag{1}$$

$$\boldsymbol{q}(\boldsymbol{x}) := \boldsymbol{x}\boldsymbol{W}^Q + \boldsymbol{b}^Q, \quad \boldsymbol{k}(\boldsymbol{x}) := \boldsymbol{x}\boldsymbol{W}^K + \boldsymbol{b}^K \tag{2}$$

ここで $\boldsymbol{x}^m, \boldsymbol{x}^n \in \mathbb{R}^d$ はクエリとキーに対応するベクトル，$\boldsymbol{W}^Q, \boldsymbol{W}^K \in \mathbb{R}^{d\times d'}$ はこのヘッドにおけるクエリとキーの重み行列，$\boldsymbol{b}^Q, \boldsymbol{b}^K \in \mathbb{R}^{d'}$ はこのヘッドにおけるクエリとキーのバイアスである．

次に，式 (1) の $\alpha_{m,n}$ における softmax 関数の引数の分子を $S = \boldsymbol{q}(\boldsymbol{x}^m)\boldsymbol{k}(\boldsymbol{x}^n)^\top$ とおき，$\boldsymbol{W} = \boldsymbol{W}^Q\boldsymbol{W}^{K\top}$ とすると S は以下のように 4 つの項の和となる．

$$S = \boldsymbol{x}^m\boldsymbol{W}\boldsymbol{x}^{n\top} + \boldsymbol{b}^Q\boldsymbol{W}^{K\top}\boldsymbol{x}^{n\top} + \boldsymbol{x}^m\boldsymbol{W}^Q\boldsymbol{b}^{K\top} + \boldsymbol{b}^Q\boldsymbol{b}^{K\top} \tag{3}$$

ここで，アテンション重みの大きさを決めるのは，同一クエリに対する異なるキー間でのスコアの差であることに留意し，式 (3) の第 1 項と第 2 項におけるキーベクトル $\boldsymbol{x}^{n\top}$ の j 次元目からの寄与 C_{1j}，C_{2j} を考える：

$$C_{1j} = \left(\sum_{i=1}^{d} x_i^m w_{ij}\right)x_j^n, \quad C_{2j} = \left(\sum_{i=1}^{d'} b_i^Q w_{ij}^k\right)x_j^n \tag{4}$$

このとき式 (3) の第 1 項，第 2 項は以下のように分解できる．

$$\boldsymbol{x}^m\boldsymbol{W}\boldsymbol{x}^{n\top} = \sum_{i,j=1}^{d} x_i^m w_{ij} x_j^n = \sum_{j=1}^{d} C_{1j} \tag{5}$$

$$\boldsymbol{b}^Q\boldsymbol{W}\boldsymbol{x}^{n\top} = \sum_{j=1}^{d}\sum_{i=1}^{d'} b_i^Q w_{ij}^k x_j^n = \sum_{j=1}^{d} C_{2j} \tag{6}$$

C_{1j} と C_{2j} の和を C_j とし，全ての次元についてまとめたベクトル $\boldsymbol{C} = [C_j]$ は以下のように計算できる．

$$\boldsymbol{C} = \left(\boldsymbol{x}^m\boldsymbol{W} + \boldsymbol{b}^Q\boldsymbol{W}^{K\top}\right) \odot \boldsymbol{x}^n \tag{7}$$

$\boldsymbol{C}$ の各成分を比較することで，アテンションスコアにおける各次元からの寄与がわかる．

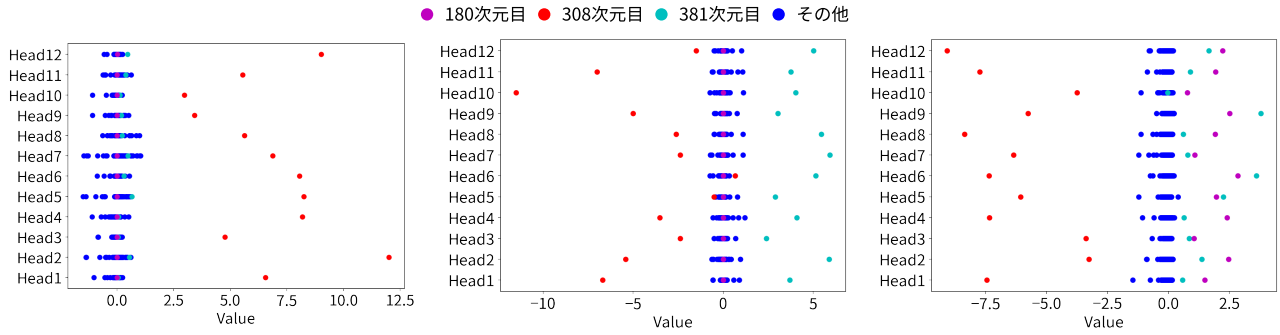
4.2 実験の概要

モデルとして BERT-base（uncased）を，入力データとして STS-B dev データセットの sentence 1 から 100 文ランダムにサンプリングしたものを用いた．BERT-base（uncased）は全 12 層にそれぞれ 12 個のヘッドが存在する．これらのうち，第 2，6，12 層について，全ての次元の C_j の入力データ全文に対する平均値を求め，各層ごとにプロットしたものを図 4 に示す．また，外れ値次元の C_j の値をキートークンの種類（[CLS]，[SEP]，カンマ・ピリオド，その他の 4 種類）ごとに平均したものを表 1 に示す．なお，図 4 に示したもの以外の層についての C_j のプロットは Appendix に掲載した．

4.3 結果

前半層　図 4a を見ると，全てのヘッドで 308 次元目の影響が正の方向に大きく働いていることがわかる．しかし，表 1 を見ると，[CLS]，[SEP]，カンマ・ピリオドの間では C_{308} の平均値 に大きな差はないことが分かる．従って，外れ値次元の影響のみではなく，他の複数次元の影響が重なることで [CLS] にアテンションが集まっていると考えられる．

中間層　図 4b を見ると，ほぼ全てのヘッドで 308 次元目の影響が負の方向に大きく働き，全てのヘッドで 381 次元目の影響が正の方向に大きく働いていることがわかる．表 1 を見ると，キートークンが [SEP] である場合，308 次元の負の影響が無く，381



(a) Layer 2　(b) Layer 6　(c) Layer 12

図 4: 層ごとの全ヘッドにおける，アテンションスコアへの各次元からの寄与の分布

表 1: キートークンの種類ごとの，外れ値次元における C_j の平均値

	token	Head1	Head2	Head3	Head4	Head5	Head6	Head7	Head8	Head9	Head10	Head11	Head12
Layer 2, 308 次元目	[CLS]	11.9	21.9	8.8	15.6	15.3	15.1	12.5	10.7	6.2	5.4	10.3	16.6
	[SEP]	10.0	18.4	7.4	13.1	12.9	12.7	10.5	9.0	5.2	4.6	8.6	14.0
	. or ,	10.7	19.6	7.7	13.1	13.4	13.0	11.2	9.1	5.6	4.8	9.1	14.6
	Other	5.6	10.2	4.0	6.9	7.0	6.8	5.9	4.8	2.9	2.5	4.7	7.7
Layer 6, 308 次元目	[CLS]	-9.7	-8.1	-3.5	-5.3	-0.8	0.9	-3.5	-3.9	-7.2	-16.8	-10.2	-2.2
	[SEP]	0.1	0.1	0.0	0.1	0.0	0.0	0.0	0.0	0.1	0.2	0.1	0.0
	. or ,	-9.6	-7.8	-3.4	-4.9	-0.8	1.0	-3.5	-3.8	-7.2	-16.2	-9.9	-2.3
	Other	-6.7	-5.4	-2.4	-3.5	-0.5	0.7	-2.3	-2.6	-5.0	-11.5	-7.0	-1.5
Layer 6, 381 次元目	[CLS]	5.1	8.0	3.2	5.6	4.0	7.0	8.1	7.5	4.1	5.5	5.1	6.8
	[SEP]	28.1	44.7	18.0	31.1	22.0	38.8	44.9	41.6	23.0	30.5	28.5	38.0
	. or ,	3.9	6.3	2.6	4.3	3.1	5.5	6.4	5.7	3.2	4.3	4.0	5.4
	Other	2.0	3.2	1.3	2.2	1.6	2.8	3.2	2.9	1.6	2.2	2.0	2.7
Layer 12, 308 次元目	[CLS]	-6.2	-2.7	-2.8	-6.0	-4.9	-6.0	-5.1	-6.8	-4.7	-3.0	-6.4	-7.3
	[SEP]	0.2	0.1	0.1	0.2	0.2	0.2	0.2	0.2	0.2	0.1	0.2	0.3
	. or ,	-3.6	-1.6	-1.5	-3.5	-2.8	-3.5	-3.0	-3.9	-2.8	-1.8	-3.7	-4.2
	Other	-8.4	-3.7	-3.8	-8.3	-6.9	-8.3	-7.2	-9.4	-6.5	-4.2	-8.7	-10.3
Layer 12, 381 次元目	[CLS]	1.0	2.4	1.5	1.2	4.0	6.3	1.4	0.1	6.7	0.1	1.6	2.9
	[SEP]	0.3	0.7	0.5	0.3	1.1	2.0	0.4	0.4	2.2	-0.1	0.4	1.0
	. or ,	0.1	0.3	0.2	0.2	0.6	0.9	0.2	0.1	0.9	0.0	0.2	0.3
	Other	0.6	1.4	0.9	0.7	2.4	3.8	0.8	0.6	4.0	0.0	0.9	1.8
Layer 12, 180 次元目	[CLS]	0.2	0.3	0.1	0.3	0.3	0.4	0.1	0.2	0.3	0.1	0.3	0.3
	[SEP]	8.0	12.2	5.5	12.5	10.1	14.7	5.4	9.7	13.2	3.8	9.8	11.2
	. or ,	6.7	11.5	4.9	10.9	9.1	13.1	4.9	8.9	11.5	3.6	8.9	10.2
	Other	0.7	1.1	0.5	1.1	0.9	1.2	0.5	0.8	1.1	0.3	0.9	1.0

次元目の正の影響が他の種類のトークンと比べて非常に大きいことがわかる．従って外れ値次元の影響で [SEP] にアテンションが集まっていると言える．

後半層　図 4c を見ると，全てのヘッドで 308 次元目の影響が負の方向に大きく働き，ほぼ全てのヘッドで 180 次元目と 381 次元目の影響が正の方向に大きく働いていることがわかる．表 1 を見ると，381 次元目の値はどの種類のトークンもほぼ同程度だが，キートークンが [SEP] である場合，308 次元目の負の影響が無いことがわかる．また，キートークンが [SEP] あるいはカンマ・ピリオドである場合，180 次元目の正の影響が他の種類のトークンと比べて非常に大きいことがわかる．以上から Layer 12 では外れ値次元の影響で [SEP]，カンマ・ピリオドにアテンションが集まっていると言える．

5　まとめ

BERT の中間層において [SEP]，最終層において [SEP] およびカンマ・ピリオドに対しアテンションが集中するのは，外れ値次元の影響が非常に大きいことがわかった．一方，前半層において [CLS] にアテンションが集中する現象は，外れ値次元の影響のみでは説明できなかった．これについては今後の課題としたい．

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 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.

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, **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, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[3] Jesse Vig. A multiscale visualization of attention in the transformer model. In Marta R. Costa-jussà and Enrique Alfonseca, editors, **Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations**, pp. 37–42, Florence, Italy, July 2019. Association for Computational Linguistics.

[4] Benjamin Hoover, Hendrik Strobelt, and Sebastian Gehrmann. exBERT: A Visual Analysis Tool to Explore Learned Representations in Transformer Models. In Asli Celikyilmaz and Tsung-Hsien Wen, editors, **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations**, pp. 187–196, Online, July 2020. Association for Computational Linguistics.

[5] Olga Kovaleva, Alexey Romanov, Anna Rogers, and Anna Rumshisky. Revealing the dark secrets of BERT. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, **Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)**, pp. 4365–4374, Hong Kong, China, November 2019. Association for Computational Linguistics.

[6] Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. What does BERT look at? an analysis of BERT's attention. In Tal Linzen, Grzegorz Chrupa la, Yonatan Belinkov, and Dieuwke Hupkes, editors, **Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP**, pp. 276–286, Florence, Italy, August 2019. Association for Computational Linguistics.

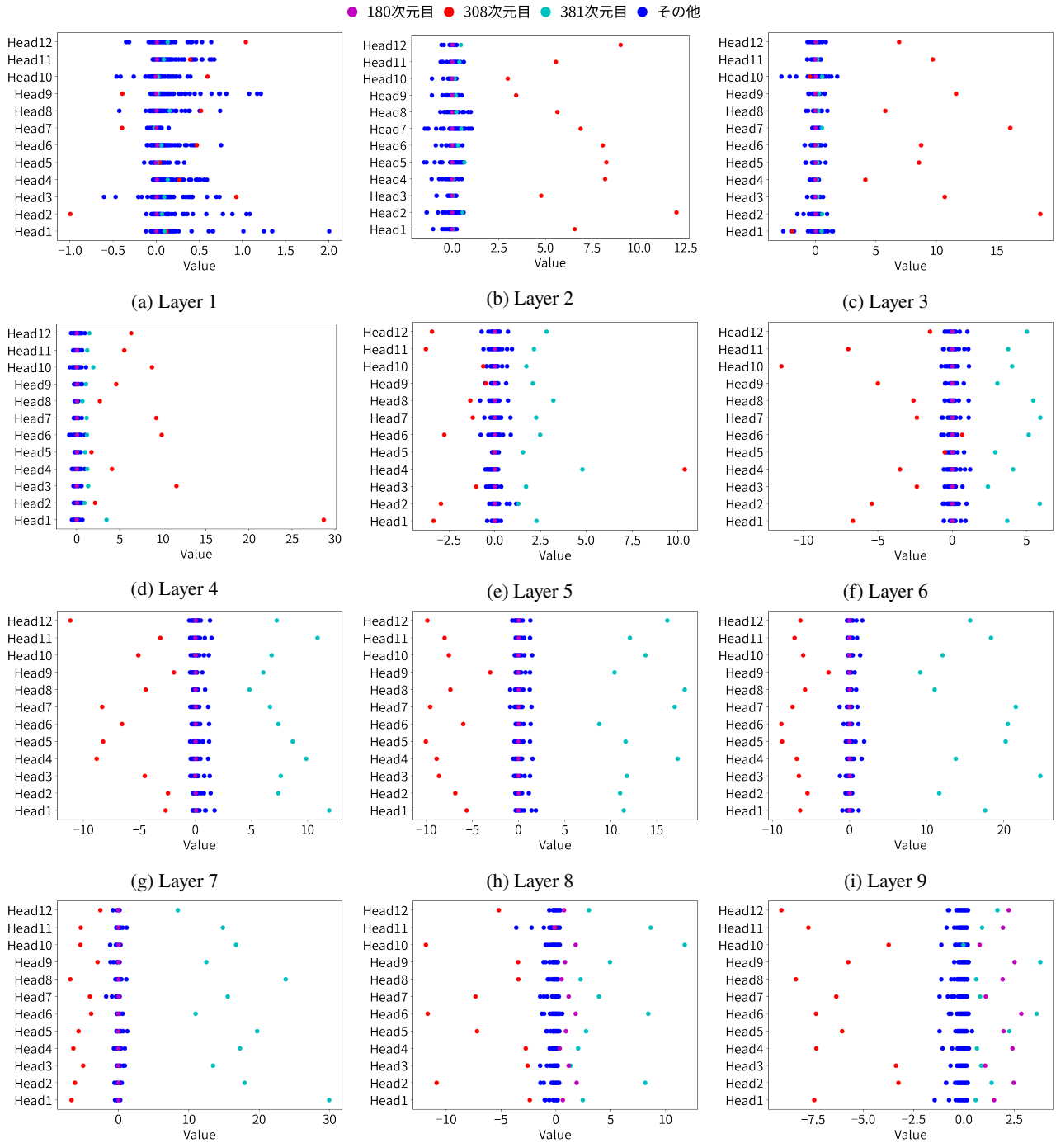
[7] Goro Kobayashi, Tatsuki Kuribayashi, Sho Yokoi, and Kentaro Inui. Attention is not only a weight: Analyzing transformers with vector norms. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)**, pp. 7057–7075, Online, November 2020. Association for Computational Linguistics.

[8] Giovanni Puccetti, Anna Rogers, Aleksandr Drozd, and Felice Dell'Orletta. Outlier dimensions that disrupt transformers are driven by frequency. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, **Findings of the Association for Computational Linguistics: EMNLP 2022**, pp. 1286–1304, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.

[9] William Rudman, Catherine Chen, and Carsten Eickhoff. Outlier dimensions encode task specific knowledge. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, **Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing**, pp. 14596–14605, Singapore, December 2023. Association for Computational Linguistics.

[10] Ziyang Luo, Artur Kulmizev, and Xiaoxi Mao. Positional artefacts propagate through masked language model embeddings. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, **Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)**, pp. 5312–5327, Online, August 2021. Association for Computational Linguistics.

[11] Olga Kovaleva, Saurabh Kulshreshtha, Anna Rogers, and Anna Rumshisky. BERT busters: Outlier dimensions that disrupt transformers. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, **Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021**, pp. 3392–3405, Online, August 2021. Association for Computational Linguistics.



(a) Layer 1　(b) Layer 2　(c) Layer 3

(d) Layer 4　(e) Layer 5　(f) Layer 6

(g) Layer 7　(h) Layer 8　(i) Layer 9

(j) Layer 10　(k) Layer 11　(l) Layer 12

図 5: 全 12 層の層ごとの全ヘッドにおける，アテンションスコアへの各次元からの寄与の分布

A　全 12 層におけるアテンションスコアへの次元ごとの寄与の分布

本文では扱えなかった層も含めた全 12 層の C_j の分布図を掲載する．Layer 1 においてはまだ支配的な次元は存在していないと言える．Layer 2 と Layer 3 では 308 次元目の影響が正の方向に大きい．Layer 4 から 381 次元目の影響が正の方向に大きくなり始め，Layer 5 では 381 次元目は正の方向，308 次元目が負の方向にはたらく．その傾向が Layer 6~Layer 10 まで続いた後，Layer 11 から 180 次元目の影響が正の方向にはたらき始め，Layer 12 では 180 次元目・381 次元目が正の方向に，308 次元目が負の方向にはたらくことがわかる．