ベクトル長に基づく自己注意機構の解析

小林 悟郎 [1]　栗林 樹生 [1,2]　横井 祥 [1,3]　鈴木 潤 [1,3]　乾 健太郎 [1,3]
[1] 東北大学　[2]Langsmith 株式会社　[3] 理化学研究所
{goro.koba,kuribayashi,yokoi,jun.suzuki,inui}@ecei.tohoku.ac.jp

1　はじめに

BERT をはじめ**自己注意機構 (self-attention mechanism)**を基盤とした深層ニューラルネットが幅広い NLP タスクで成功を収めている [4, 10, 11, 13]．とくに近年は，モデルの成功の理由を理解するため，またモデルが捉える言語現象を確かめるため，自己注意機構自体の分析が盛んに取り組まれている [2, 8, 9]．自己注意機構は入力ベクトル列の中から特に必要なベクトルに注目し，ここから情報を集めて新たなベクトルを作る役割を担う．具体的には，入力系列の各入力ベクトルをアフィン変換した上で重み付け和をし出力ベクトルを作る（図1）．自己注意機構の分析における主たる関心事も「ある出力ベクトルを作るためにどの入力ベクトルが注目されているか」であり，これを調べるために各入力に割り振られる**アテンション重み**（図1，重み）の大小が確かめられてきた [2, 3, 8, 9, 12]．

アテンション重みの大小の分析は直感的ではあるが，ここでは入力ベクトルの大きさや自己注意機構におけるアフィン変換の影響（図1，変換後のベクトルの大きさ）は考慮されない．結果として，アテンション重みの大小と実際に足し合わされるベクトル（図1，足し合わされるベクトル）の大小には乖離が生まれ得る．言い換えれば，「足し合わせ」をおこなう機構がどの入力を重く足し合わせているかについて分析したいにも関わらず，実際に足し合わされているベクトルの大きさとは全く異なる量が分析されている可能性がある（4節）．例えば図 1中の $\boldsymbol{y}_4$ を作る際，従来通りアテンション重みの観察に基づくと $\boldsymbol{x}_1$ に大きく注目が集まっているように見えるが，実際に完成するベクトルには $\boldsymbol{x}_1$ はほとんど寄与しておらず，もっとも寄与しているのは $\boldsymbol{x}_3$ である．

上述の問題に対処するため，変換後のベクトルと重みの両者を考慮した**足し合わされるベクトルの大きさ**によって各入力ベクトルが出力ベクトルに寄与する度合いを測ることを提案する．実験セクションでは，自己注意機構を基盤としたモデルとして BERT[4] に焦点を当て，自己注意機構全体の分析を行う．実験から，自己注意機構におけるアフィン変換が入力ベクトルの特定の情報を拡大・縮小する作用を持つこと，重みが同じでも変換後のベクトルの大きさに分散があることを示す．これらの結果より，自己注意機構における，アテンション重みからは推定できない作用の存在が示唆される．

次に，ベクトルの大きさに注目した分析結果と従来のアテンション重みに基づく分析結果の差異について分析する．アテンション重みによる分析結果では，出力ベクトルに対する特殊トークンの寄与が大きいという直感に反する傾向が見られていたが，ベクトルの大きさに注目した分析では，特殊トークンの寄与は小さかった．このことから，特殊トークンはある種アテンションの「ゴミ箱」のような役割をしていることが示唆された．

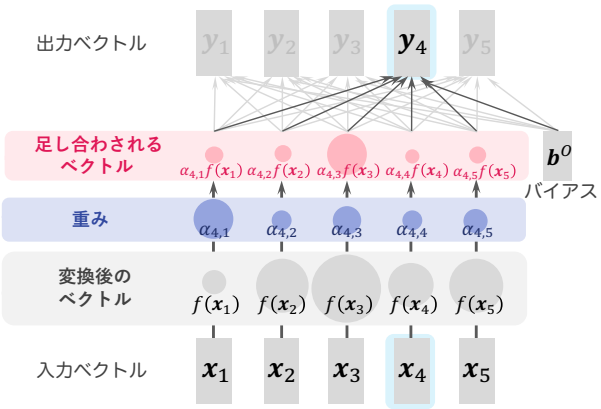


図1: 自己注意機構の概要．変換・重みづけしたベクトルを足し合わせることで出力ベクトルを作る．ベクトルに対応した円の大きさはベクトル長を指す．

本研究の貢献は以下の通りである．

- 自己注意機構の挙動について，これまで考慮されてこなかった変換後のベクトルを考慮する分析方法を提案した．
- BERT の自己注意機構を分析し，[CLS] などの特殊トークンではアテンション重みと変換後のベクトルの大きさが打ち消しあう傾向にあることを示した．

2　関連研究

多くの既存研究では，自己注意機構を基盤としたモデルとして，BERT などの事前学習言語モデルを分析の対象としている．事前学習言語モデルの分析アプローチは大きく三つある．

一つは，モデルの埋め込みや中間表現から，対象となる言語現象の情報を復元可能かを検証する試みである．この方法を用いた先行研究では，例えば，BERT の中間表現から統語構造や意味情報が高い精度で復元可能であることから，BERT はそれらの言語的性質を捉えていると報告されている [3, 5, 7]．

もう一つは，入力ベクトルやアテンション重みが出力ベクトルや損失関数に与える影響を勾配によって測る試みである．この方法を用いた先行研究では, BERT の中間層において, [SEP] に対応するアテンション重みは，穴埋め言語モデルの損失関数にあまり影響を及ぼさないことなどが報告されている [1, 2]．

自己注意機構が割り振る重みに着目する分析も，典型的なアプローチの一つである．この方法を用いた先行研究では，BERT は特殊トークンに対応するベクトルに大きな重みを割り振る傾向があることや，BERT および RoBERTa には特定の依存関係に対応する部分に大きな重みを割り振るヘッドが存在することが報告されている [2, 8, 9]．本研究では，割り振られた重みから自己注意機構の挙動を分析する方法に対する指摘と新たな分析方法を提示する．

3 準備：自己注意機構

3.1 自己注意機構の概略

自己注意機構は入力ベクトルの組 $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\} \subseteq \mathbb{R}^d$ から，各出力ベクトル $\boldsymbol{y}_i \in \mathbb{R}^d$ を次のように計算する[*1]：

$$\boldsymbol{y}_i = \left(\sum_{j=1}^{n} \alpha_{i,j} \boldsymbol{v}(\boldsymbol{x}_j) \right) \boldsymbol{W}^O + \boldsymbol{b}^O \in \mathbb{R}^d \tag{1}$$

$$\alpha_{i,j} := \underset{j}{\mathrm{softmax}} \left(\frac{\boldsymbol{q}(\boldsymbol{x}_i) \boldsymbol{k}(\boldsymbol{x}_j)^\top}{\sqrt{d'}} \right) \in \mathbb{R} \tag{2}$$

$$\boldsymbol{q}(\boldsymbol{x}) := \boldsymbol{x}\boldsymbol{W}^Q + \boldsymbol{b}^Q \quad \left(\boldsymbol{W}^Q \in \mathbb{R}^{d \times d'}, \boldsymbol{b}^Q \in \mathbb{R}^{d'}\right) \tag{3}$$

$$\boldsymbol{k}(\boldsymbol{x}) := \boldsymbol{x}\boldsymbol{W}^K + \boldsymbol{b}^K \quad \left(\boldsymbol{W}^K \in \mathbb{R}^{d \times d'}, \boldsymbol{b}^K \in \mathbb{R}^{d'}\right) \tag{4}$$

$$\boldsymbol{v}(\boldsymbol{x}) := \boldsymbol{x}\boldsymbol{W}^V + \boldsymbol{b}^V \quad \left(\boldsymbol{W}^V \in \mathbb{R}^{d \times d'}, \boldsymbol{b}^V \in \mathbb{R}^{d'}\right) \tag{5}$$

Query ベクトル $\boldsymbol{q}(\boldsymbol{x}_i)$ と Key ベクトル $\boldsymbol{k}(\boldsymbol{x}_j)$ から重み $\alpha_{i,j}$ を計算し，Value ベクトル $\boldsymbol{v}(\boldsymbol{x}_j)$ を重み付け和し，最後に $\boldsymbol{W}^O \in \mathbb{R}^{d' \times d}$，$\boldsymbol{b}^O \in \mathbb{R}^d$ によるアフィン変換を加える．なお，Transformer および BERT などの事前学習言語モデルでは，この自己注意機構（“ヘッド”）を並列に並べ，各ヘッドの出力ベクトルを足し合わせるマルチヘッド自己注意機構が使われている．

3.2 自己注意機構は変換したベクトルを足し合わせる機構

式1は，行列積 $\boldsymbol{W}^O$ の線形性から，次のように変形できる．

$$\boldsymbol{y}_i = \sum_{j=1}^{n} \alpha_{i,j} \boldsymbol{v}(\boldsymbol{x}_j) \boldsymbol{W}^O + \boldsymbol{b}^O \tag{6}$$

$$= \sum_{j=1}^{n} \alpha_{i,j} \, f(\boldsymbol{x}_j) + \boldsymbol{b}^O \tag{7}$$

$$f(\boldsymbol{x}) = (\boldsymbol{x}\boldsymbol{W}^V + \boldsymbol{b}^V)\, \boldsymbol{W}^O \tag{8}$$

図1の自己注意機構の概要は式 7に基づく．式7より，自己注意機構は各入力ベクトル $\boldsymbol{x}$ を f で変換して $f(\boldsymbol{x})$ を作った後，計算した重み α をかけた $\alpha f(\boldsymbol{x})$ をバイアス $\boldsymbol{b}^O$ と共に足し合わせる機構とみなせる．

4 提案手法：足されるベクトルの大きさの分析

自己注意機構を分析する先行研究では，「各入力にどれほど重みが割り振られたか」を見ることで機構がどの入力を重く足し合わせているかを分析できるという暗黙の仮定をおき，重み α の大きさを分析している [2, 3, 8, 9, 12]．式 7より，この手法では変換後のベクトル $f(\boldsymbol{x})$ の影響が無視されている．本稿では，実際に足し合わされるベクトル $\alpha f(\boldsymbol{x})$ の大きさを L2 ノルムで計算することで，重み α に加えて変換後のベクトル $f(\boldsymbol{x})$ の影響も考慮して，各入力ベクトルが出力ベクトルに寄与する度合いを測る分析方法を提案する．5.2節では，$f(\boldsymbol{x})$ の作用について分析し，この作用が入力の足し合わせ具合に大きな影響をもち得ることを示す．

[*1] なお，本稿ではベクトルは横ベクトルとしている．

表1: BERT における変換後のベクトルの大きさ $\|f(\boldsymbol{x})\|$ のばらつきの度合い

ヘッド	平均	標準偏差	変動係数	最大値	最小値
2 層 4 ヘッド（変動係数最大）	4.26	1.59	**0.37**	12.66	0.96
2 層 7 ヘッド（変動係数最小）	3.40	0.50	**0.12**	6.15	1.35
全ヘッド	5.15	1.75	**0.34**	18.64	0.10

5 実験

アテンション重みを測る既存手法と提案手法を用いて BERT の自己注意機構を分析する．5.2節では既存手法が無視してきた影響について，5.3節では重み と 変換後のベクトル の関係について調査する．5.4節では，重み と 足し合わされるベクトル の関係を調査し，これまでの検証で得られていた知見を再検証する．

5.1 実験設定

■ **モデル** 解析が最も盛んな事前学習済みの BERT-base (uncased) [*2]を用いた．BERT-base は 12 層から構成され，各層は 12 個の自己注意機構（ヘッド）を持つため，モデル全体では 144 個の自己注意機構を持つ．

■ **データ** BERT の挙動を分析するためのテキストデータとして，先行研究 [2] が公開している 992 の入力系列を用いた．各入力系列には，英語 Wikipedia から抽出した二つの連続したパラグラフが以下の形式で格納されている：`[CLS] paragraph1 [SEP] paragraph2 [SEP]`．系列に含まれるトークン数は，最小 18，最大 128，平均 122.11 である．

5.2 重みを分析する既存手法が無視するもの

式 8の $f(\boldsymbol{x})$ の大きさ（以降，$\|f(\boldsymbol{x})\|$）がおおよそ一定なら，重み α は実際に足し合わされるベクトル $\alpha f(\boldsymbol{x})$ の大きさ（以降，$\|\alpha f(\boldsymbol{x})\|$）の良い近似となる．しかし，もし $\|f(\boldsymbol{x})\|$ が大小様々な値を取るなら，重み α と実際に足し合わされるベクトルの大きさ $\|\alpha f(\boldsymbol{x})\|$ の間には大きな乖離が生まれることになる．

■ **$\|f(\boldsymbol{x})\|$ のばらつきの度合い** $\|f(\boldsymbol{x})\|$ のばらつき具合を調べるため，$\|f(\boldsymbol{x})\|$ の変動係数（$\frac{\text{標準偏差}}{\text{平均}}$）の大きさを調査した．事前学習済みの BERT-base に 992 系列を入力し，各ヘッドにおける $\|f(\boldsymbol{x})\|$ の変動係数を求めた．変動係数が最大となるヘッド，最小のヘッドにおける変動係数と，全ヘッドにおける変動係数を表1に示す．表1より，$\|f(\boldsymbol{x})\|$ には変動係数で 0.34 (ヘッド平均) のばらつきがある．これは，$\|f(\boldsymbol{x})\|$ の値が平均のおよそ 0.66 倍から 1.34 倍の値を容易にとり得ることを意味し，無視できないばらつきであることが示唆される．このことは，$\|f(\boldsymbol{x})\|$ の最大値と最小値の比較からも見てとれる (表1)．以降の段落では，$\|f(\boldsymbol{x})\|$ のばらつきの要因を探るため，(i) $\|\boldsymbol{x}\|$ のばらつきの度合いと，(ii) f による拡大・縮小作用に焦点を当てて分析を行う．

■ **$\|\boldsymbol{x}\|$ のばらつきの度合い** $\|\boldsymbol{x}\|$ のばらつきの度合いを変動係数の大きさから確かめる．実験設定は，$\|f(\boldsymbol{x})\|$ の変動係数の分析時と同様である．$\|\boldsymbol{x}\|$ の変動係数を表2に示す．$\|\boldsymbol{x}\|$ にもばらつきが存在するものの，$\|f(\boldsymbol{x})\|$ の変動係数 (表3) と比較すると，$\|\boldsymbol{x}\|$ のばらつきは比較的小さいことが分かる．このことから，f に拡大・縮小作用があることが示唆される．

[*2] https://github.com/huggingface/transformers の実装を用いた．

表2: BERT における入力ベクトルの大きさ $\|\boldsymbol{x}\|$ のばらつきの度合い

層	平均	標準偏差	変動係数	最大値	最小値
12（変動係数最大）	20.49	4.62	**0.23**	32.84	4.13
7（変動係数最小）	21.64	1.40	**0.06**	23.03	11.87
全層	19.93	3.03	**0.15**	32.84	4.13

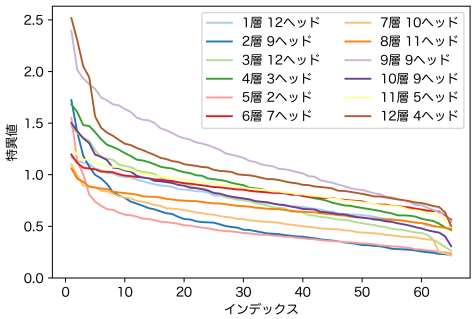


図2: BERT-base のランダムな自己注意機構における変換 f の特異値

■ f による拡大・縮小作用　関数 f が入力 $\boldsymbol{x}$ をどの程度拡大・縮小するかを確認する．一般にアフィン変換 $f\colon \mathbb{R}^d \to \mathbb{R}^d$ は線形変換 $\mathbb{R}^{d+1} \to \mathbb{R}^{d+1}$ と同一視できるので[*3]，その拡大・縮小の作用は線形変換の特異値を見れば確かめることができる [6]．実際に事前学習済みの BERT-base から自己注意機構（ヘッド）をランダムに選択し[*4]，各ヘッドにおける変換 f の特異値を降順に並べた結果を図2に示す．いずれのヘッドにおいても，f によって入力が最も拡大される場合と最も縮小される場合とでは，拡大率に少なくとも 1.8 倍以上の差があることが分かる．すなわち，BERT の内部では変換 f によって $\|f(\boldsymbol{x})\|$ の値に大きなばらつきが生まれている．このことから，重み α と実際に足し合わされるベクトルの大きさ $\|\alpha f(\boldsymbol{x})\|$ の間には大きな乖離があることが分かる．

5.3　実験 2: アテンション重み α と $\|f(x)\|$ の関係

この節では，アテンション重み α（既存手法）と $\|f(\boldsymbol{x})\|$（既存手法が無視している部分）の関係を確かめる．

結果，同じ程度の α に対して $\|f(\boldsymbol{x})\|$ は大きくばらつくこと，また一部のトークンでは α と $\|f(\boldsymbol{x})\|$ の大小が逆転する傾向にあることが確かめられた．すなわち，α を見るだけでは実際に足し合わされるベクトルの大きさ $\|\alpha f(\boldsymbol{x})\|$ は推察できないと考えられる．

■ 実験設定　全系列をモデルに入力し，モデル内の全自己注意機構で α と $\|f(\boldsymbol{x})\|$ 確認する．

■ 結果　α と $\|f(\boldsymbol{x})\|$ の組をトークン種別ごとに 1% ずつ乱択し図3に示した．図3より，同じ重み α が割り振られていても変換後のベクトルの大きさ $\|f(\boldsymbol{x})\|$ にはばらつきがある（大きな値も小さな値も取りうる）ことがわかる．すなわち，重み α を見るだけでは，自己注意機構が実際に足し合わせるベクトルの大きさ $\|\alpha f(\boldsymbol{x})\|$ を推定することは難しい．

[*3] 各ベクトル $\boldsymbol{x}$ の末尾に 1 を加えて $\tilde{\boldsymbol{x}} := \begin{bmatrix} \boldsymbol{x} & 1 \end{bmatrix} \in \mathbb{R}^{d+1}$ とすると，アフィン変換 f は次の線形変換 $\tilde{f}$ と同一視できる．

$$\tilde{f}(\tilde{\boldsymbol{x}}) = \tilde{\boldsymbol{x}}\widetilde{\boldsymbol{W}}^V\widetilde{\boldsymbol{W}}^O \tag{9}$$

$$\widetilde{\boldsymbol{W}}^V := \begin{bmatrix} \boldsymbol{W}^V & \begin{matrix}0\\\vdots\\0\end{matrix} \\ \boldsymbol{b}^V & 1 \end{bmatrix} \in \mathbb{R}^{(d+1)\times(d'+1)} \tag{10}$$

$$\widetilde{\boldsymbol{W}}^O := \begin{bmatrix} \boldsymbol{W}^O & \begin{matrix}0\\\vdots\\0\end{matrix} \\ \begin{matrix}0 & \dots & 0\end{matrix} & 1 \end{bmatrix} \in \mathbb{R}^{(d'+1)\times(d+1)} \tag{11}$$

[*4] 12 層それぞれから各 1 ヘッドをランダムに選択した．

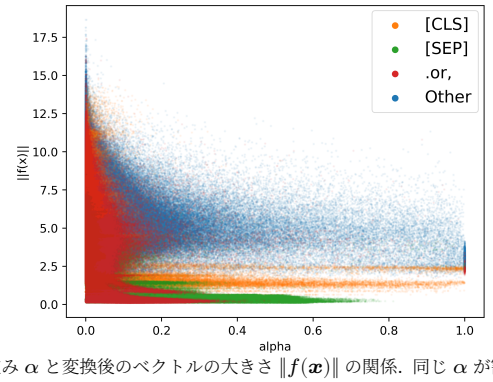


図3: 重み α と変換後のベクトルの大きさ $\|f(\boldsymbol{x})\|$ の関係．同じ α が割り振られていても $\|f(\boldsymbol{x})\|$ には分散があることがわかる．

表3: α と $||f(\boldsymbol{x})||$ のベクトルの種類ごとのスピアマンの順位相関係数

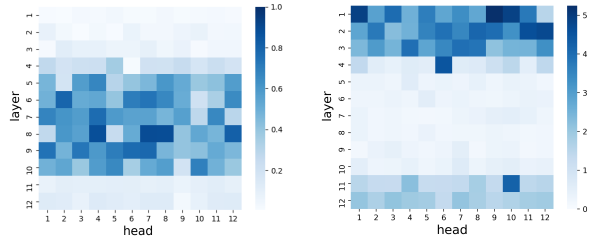
種類	ベクトル数	スピアマン
[CLS]	17,443,296	-0.34
[SEP]	34,886,592	-0.69
カンマ・ピリオド	182,838,528	-0.25
その他	1,944,928,224	-0.06

加えて，全体として α と $\|f(\boldsymbol{x})\|$ には反比例の関係が見て取れる．トークン種別ごとの両者の順位相関係数を表3に示す．特殊トークンである [CLS] および [SEP] とカンマ・ピリオドに対応するベクトルでは負の相関が高いことがわかる．たしかに図3からも，これらの特殊トークンは比較的大きな α を持つが $\|f(\boldsymbol{x})\|$ が小さいことが見て取れる．これらの特殊トークンの果たす役割については次の節で詳しく述べる．

次に，α と $\|f(x)\|$ がとりわけ強い負の相関を持つ [SEP] トークンについて詳細に分析した．図4aは 144 個の自己注意機構それぞれにおいて，系列内の全ての [SEP] トークンに対応するベクトルに割り振られる重み α の合計値を全系列で平均したものである．図4bは $\|f(\boldsymbol{x})\|$ の値を，入力データに含まれる全ての [SEP] トークンに対応するベクトルに関して平均したものである．両図では，縦が層を表し，横にその層内の自己注意機構が並んでいる．各マスでは色が濃いほど値が大きいことを表す．図4より，[SEP] トークンに対応するベクトルでは，割り振られる重み α と変換後のベクトルの大きさ $\|f(\boldsymbol{x})\|$ の大小関係が逆転していた．また，α と $\|f(\boldsymbol{x})\|$ のどちらか一方が非常に小さい値を取る傾向にあることがわかる．[CLS] およびピリオド・カンマに対応するトークンでも，同様の傾向が見られた．その他のベクトルでは，強い傾向は見られなかった．これより，これらのトークンに対応するベクトルでは，α と $\|f(\boldsymbol{x})\|$ の片方がもう片方を打ち消してしまうため，足し合わされるベクトルの大きさ $\|\alpha f(\boldsymbol{x})\|$ は小さくなりやすいと言える．

5.4　実験 3: アテンション重み α と $\|\alpha f(x)\|$ の関係

5.3節で述べたように，BERT の自己注意機構は特定のトークンに対応するベクトルにおいて，重み α と $\|f(\boldsymbol{x})\|$ が打ち消し合う傾向がある．この節では，重み α がこれらのベクトルに偏る傾向があるにも関わらず，これらのベクトルの寄与 $\|\alpha f(\boldsymbol{x})\|$ は小さいことを述べる．結論として，入力ベクトル全体に集めたい情報がない場合に，ある種のゴミ箱（重みの逃げ



(a) 系列内で [SEP] トークンに対応するベクトルに割り振られる α の合計値の平均

(b) [SEP] トークンに対応するベクトルの $\|f(\boldsymbol{x})\|$ の平均

図4: [SEP] トークンに対応するベクトルのヘッド毎の分析結果

先）としてこれらのベクトルが使われていることが推測された．

■ 実験設定　全 992 系列をモデルに入力し，モデル内の各自己注意機構において割り振られる重み α と足し合わされるベクトルの大きさ $\|\alpha f(\boldsymbol{x})\|$ を対応するトークンの種類ごとに取り出し，層ごとに平均した．各入力系列には，[CLS] が 1 個，[SEP] が 2 個，ピリオドとカンマが合わせて平均約 10 個，その他のトークンが平均約 109 個含まれている．クラークら [2] に従い，各系列内でそれぞれの種類のトークンに対応するベクトルに割り振られる重み α および足し合わされるベクトルの大きさ $\|\alpha f(x)\|$ をそれぞれ合計し，全系列で平均した．

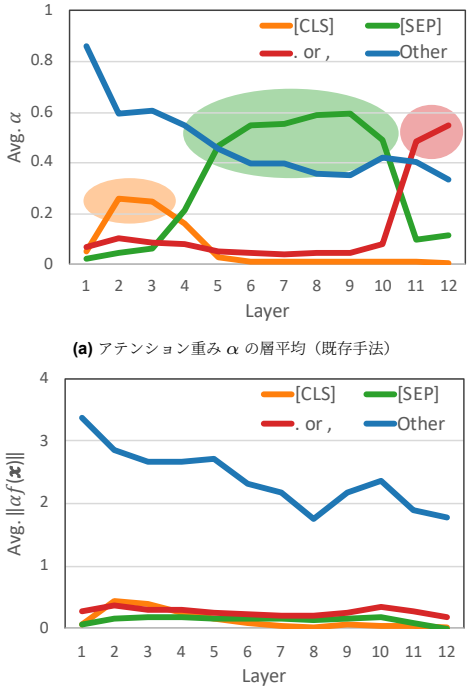
■ 結果　図5aより，BERT の自己注意機構は前半層で [CLS] に，中間層で [SEP] に，後半層でピリオドおよびカンマに対応するベクトルに大きな重み α を割り振ることがわかる（以降，[CLS]，[SEP]，ピリオド，カンマに対応するベクトルを $\boldsymbol{x}_{\text{special}}$，その他のベクトルを $\boldsymbol{x}_{\text{other}}$ とする）．特に 5 層目以降では，$\boldsymbol{x}_{\text{special}}$ に対する重み α が，その他のベクトルに割り振られる重みよりも大きく，全体の半分以上を占めることがわかる．この結果は，先行研究 [2, 9] の結果と一致する．

一方で，図5bより，ベクトル $\boldsymbol{x}_{\text{special}}$ から足されるベクトルの大きさ $\|\alpha f(\boldsymbol{x}_{\text{special}})\|$ は小さいことが分かる．$\boldsymbol{x}_{\text{other}}$ から足されるベクトルの大きさ $\|\alpha f(\boldsymbol{x}_{\text{other}})\|$ の合計の方がはるかに大きく，重みの結果とは大小関係が大きく異なっていた．これより，BERT のほとんどの層では自己注意機構が特定のトークンに対応するベクトルに大きな重みを割り振る傾向があるにも関わらず，それらのベクトルの寄与は小さいことがわかった．

式1, 2より，自己注意機構は出力ベクトルを作る際に，各入力ベクトルに対して重み α を合計が 1 になるように割り振る．したがって，入力に集めたい情報が少なく，出力ベクトル $\boldsymbol{y}_i$ を小さくしたい場合には，変換後のベクトルの大きさ $\|f(\boldsymbol{x})\|$ が小さい入力ベクトル $\boldsymbol{x}$ に重み α を割り振る必要がある．BERT では，[CLS]，[SEP]，ピリオド・カンマといった，ほとんど全ての入力系列で出現し，意味を多く持たないようなトークンに対応するベクトルが小さい $\|f(\boldsymbol{x})\|$ をもち，ある種のアテンションのゴミ箱のような役割を担っていたことが推測される．

6　おわりに

本稿では，自己注意機構がどの入力を重く足し合わせているかは，重み α を用いた既存手法では分析できないことを示した．そこで，実際に足し合わされるベクトルの大きさ $\|\alpha f(\boldsymbol{x})\|$ を測るという手法を提案し，提案した手法を用いて BERT の自己注意機構を分析することで，既存手法とは異なる結果と特定



(a) アテンション重み α の層平均（既存手法）

(b) 足されるベクトルの大きさ $\|f(\boldsymbol{x})\|$ の層平均（提案手法）

図5: 既存手法と提案手法の分析結果．α は特定のトークンに対応するベクトルに強く割り振られる傾向があるが，それらからの $\|f(\boldsymbol{x})\|$ は大きくないことがわかる．

のトークンに対する特殊な挙動を報告した．

今後の方向性として，既存手法を用いて行われてきた分析を提案手法により再実験し，自己注意機構ベースのモデルが捉えているとされてきた言語現象を本当に捉えているのかを確認するという方向が考えられる．また，事前学習言語モデルの自己注意機構以外の部分（フィードフォワードネットワークや残差接続など）についても解析的に分析することで，モデル全体で入力から出力にどのように情報が流れるのかを把握することも興味深いと考えられる．

謝辞 本研究は JSPS 科研費 JP19H04162 の助成を受けたものです。

参考文献

[1] G. Brunner et al. “On Identifiability in Transformers”. In: *ICLR*. 2020.
[2] K. Clark et al. “What Does BERT Look At? An Analysis of BERT’s Attention”. In: *ACL Workshop BlackboxNLP*. 2019, pp. 276–286.
[3] A. Coenen et al. “Visualizing and Measuring the Geometry of BERT”. In: *arXiv preprint arXiv:1906.02715* (2019).
[4] J. Devlin et al. “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”. In: *NAACL-HLT*. 2019, pp. 4171–4186.
[5] Y. Goldberg. “Assessing BERT’s Syntactic Abilities”. In: *arXiv preprint arXiv:1901.05287* (2019).
[6] G. H. Golub and C. F. Van Loan. *Matrix Computations*. Fourth Edition. The Johns Hopkins University Press, 2013.
[7] J. Hewitt and C. D. Manning. “A Structural Probe for Finding Syntax in Word Representations”. In: *NAACL-HLT*. 2019, pp. 4129–4138.
[8] P. M. Htut et al. “Do Attention Heads in BERT Track Syntactic Dependencies?” In: *arXiv preprint arXiv:1911.12246* (2019).
[9] O. Kovaleva et al. “Revealing the Dark Secrets of BERT”. In: *EMNLP-IJCNLP* (2019), pp. 4364–4373.
[10] Z. Lan et al. “ALBERT: A Lite BERT for Self-supervised Learning of Language Representations”. In: *ICLR*. 2020.
[11] Y. Liu et al. “RoBERTa: A Robustly Optimized BERT Pretraining Approach”. In: *arXiv preprint arXiv:1907.11692* (2019).
[12] P. Michel et al. “Are Sixteen Heads Really Better than One?” In: *NIPS*. 2019, pp. 14014–14024.
[13] Z. Yang et al. “XLNet: Generalized Autoregressive Pretraining for Language Understanding”. In: *NIPS*. 2019, pp. 1–18.