BrainBERT:脳活動とテキストの対応関係を捉えた言語モデル構築への取り組み

理学専攻・情報科学コース
2040666
羅桜

1　はじめに

深層学習の発展に伴い，その手法を用いて，テキストの意味理解，音声や動画像などのマルチメディアからの情報抽出など，様々な課題において優れた成果が得られている．そのような背景は，脳神経科学の分野にも大きな影響を与え，脳内情報処理機構の解明や脳内情報解読などにも深層学習モデルが導入され，新しい研究手段として確立されてきている．しかし，fMRI など脳活動データは取得にかかるコストが大きいことや被験者の負担が大きいことなどから実験データ数や実験目的に対する用途は限られており，それらデータを使って訓練されたモデルの精度は期待したほど高くない傾向がある．一方で，汎用言語モデル BERT [1] は，自然言語処理（NLP）の分野において顕著な成功を収めている．このことを踏まえ，本研究では，BERT が言語の意味を表象するものとして脳内情報解読にも貢献可能であるかどうかを調べることを目的とし，脳活動データから重要な特徴量を抽出し，言語の意味と合わせた汎用言語モデル BrainBERT の構築を行う．具体的なアプローチとして，データ圧縮後の歪みを最小限に抑えるために，特徴抽出には Autoencoder モデル [2] を使用し，言語モデルの訓練に脳活動の特徴量を反映できるようにするための，事前学習とファインチューニングを設定し，言語と脳活動の双方の特徴量を踏まえた汎用言語モデル BrainBERT の構築を行う．

2　脳内情報解読のための汎用言語モデルの構築

本研究では，ヒト脳に与えられた刺激に対して観測された脳活動データを通じてその内容を言語に結びつけ解読することを目指し，新しい汎用言語モデル BrainBERT を構築する．モデルを構築するための脳活動データとして，文章を読んだ際の fMRI による観測データを使用し，文章の内容と脳活動とのペアデータを用いて，その対応関係を深層学習を用いて学習し，脳活動に対する意味表象を表現するする汎用言語モデル BrainBERT を構築する．学習の全体的な流れを図 1 に示す．①言語刺激下での脳活動データを用いて特徴抽出モデルの事前学習を行う．これを元に Autoencoder を用いて脳の潜在状態を抽出する．②脳特徴量とそれに対応するテキストの BERT による特徴量を反映する潜在状態を抽出し，事前学習を通じて BrainBERT を構築する．構築された BrainBERT はタスクごとにファインチューニングが適用され使用される．

2.1　データ

Alice データセット　本研究で用いるデータセットの一つは，Bhattasali ら [3] によって提供されている「不思議の国のアリス」の第 1 章を傾聴した際の聴覚刺激に

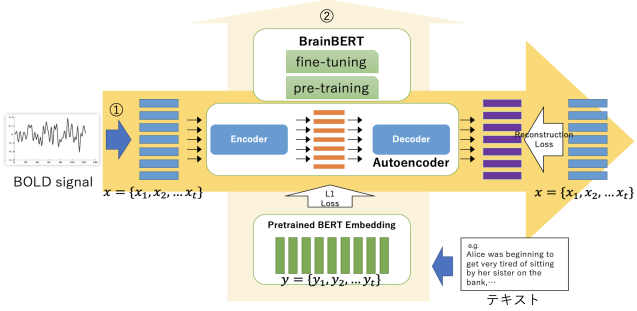

図 1: 全体的な流れ

基づく脳活動データを機能的磁気共鳴画像法（fMRI）にて収集したもの（以下，Alice データセット）である．データ収集実験において，29 人の被験者から血中酸素レベル依存性（BOLD）信号を fMRI を用いて収集した．本研究では，29 人の被験者のうち 23 人のデータを使用した．Alice データセットは元データの前処理に基づき，エッジ効果とバウンダリ効果を除外するために，最初の 10 サンプル分のデータは使用せず，10 番目の取得ポイントから合計 372 個のデータを利用することから，実際に使用したタイムポイント数は 362 とする．

Pereira データセット　本研究では，Pereira ら [4] が収集したヒト脳の画像（以下，Pereira データセット）も使用する．Pereira らはデータ収集のために 3 つの実験を行ったが，本研究では実験 2，3 におけるデータを使用した．実験 2 では，8 人の被験者に 384 個の自然言語文を視覚的に提示し，実験 3 では，6 人の被験者に他の 243 個の自然言語文を提示した．これらの文章は，Pereira らが書いた自然現象に関する簡単な百科事典のような文章で構成されている．被験者は，一文ずつその意味を考えながら丁寧に読み，文章を読んでいる際の脳活動を fMRI で記録した． 各被験者，各文ごとの fMRI データは，約 20 万次元のベクトルで構成されている．

2.2　実験 1:脳特徴量抽出における異種 Autoencoder 間の比較

異なる種類の Autoencoder がどの程度，脳の特徴量を抽出できるかを検証する．ここでは，一般的な Autoencoder（以下，Vanilla Autoencoder），Deep Autoencoder（以下，DAE），3D Convolutional Autoencoder（以下，3D-CAE）の 3 種類の Autoencoder を対象とする．

損失関数の学習曲線を観察することによって，Autoencoder の学習状況を確認する．また，検証セット

のテキストによる埋込ベクトルと抽出された脳の特徴量との相関関係をピアソン相関係数を用いて計算し，上述した 3 つのモデルの内，どのモデルがより優れた特徴量抽出能力と意味的な捕捉能力を持っているかを検証する．

損失の変化によると、Vanilla Autoencoder（BERT_{Large}）と 3D CAE モデルの方が学習の収束性が高く，これら 2 つのモデルが脳の活動データの特徴をより良く抽出できていることが分かった．

しかし，ピアソン相関係数においては，3D CAE で抽出した脳活動データは，Autoencoder モデルで抽出した脳特徴量の意味的な相関に比べて劣っていた．以上の分析の結果，Vanilla Autoencoder（BERT_{Large}）が，脳の特徴量を抽出する上で，最も優れていることが分かった．このことから，以後の実験や分析では，この学習済みモデルを使って脳の特徴を抽出する．

2.3 実験 2:NLP タスクの実験

実験 2 では，BrainBERT に上述した 2 つの学習タスクのトレーニングを中心に行う．まず，MLM タスクでは，以下に続く BERT の古典的な処理に従う．入力文をトークン化し，トークンの 15 %をランダムにマスクし，ランダムデータの 80 %を [MASK] マーク（コードは [103]）に置き換え，10 %をランダムな単語で埋め，10 %をそのままにする．トレーニングの評価基準は，ランダムマスクの位置を出力し，Softmax 関数を用いて最も予測確率の高いコードを取り出し，出力されたラベルが入力されたラベルと一致するかどうかをモデルの精度として計算する．そして，学習済みモデルは，Alice データセットのデータを使用し，5 または 10 エポック学習する．評価基準は，これまで同様，マッチング精度を採用する．

また，BTM タスク（脳活動データとテキストの照合を行うタスク）を行う．BTM への入力は，文と脳活動データとのペア集合であり，出力はその集合からサンプリングされたペアが一致するかどうかを示す二値ラベル $\boldsymbol{y} \in \{0,1\}$ となる．BERT において文の埋込ベクトルとなる [CLS] トークンの表現を入力された脳活動データとテキストのペアとの双方の埋込ベクトルから算出した表現として抽出し，それを FC 層と Sigmoid 関数に与え，0 と 1 の間のスコアを予測する．ここで，出力スコアを $s_\theta(\boldsymbol{w}, \boldsymbol{z})$ とする．訓練プロセスの各ステップで，正または負のペア $(\boldsymbol{w}, \boldsymbol{z})$ を用いて学習する．負のペアは，ペアになったサンプルの脳活動データやテキストを他のサンプルからランダムに選択されたサンプルに置き換えることで作成する．ポジティブなデータとネガティブなデータのペアの比率は約 1:1 である．BrainBERT は，事前に学習されたモデルに基づいて，MLM タスクのプレーンテキストタスクの学習において 35.46 %の精度を達成し，脳データセットの学習において 15.79 %の精度を達成した．また，BTM の結果は 53.0 %の精度を達成した．

2.4 実験 3:アテンションの追跡

脳データを追加した後の BrainBERT モデルが BERT の内部構造に与える影響をさらに理解するために，実験 3 を行った． 実験 3 では，MLM タスク下でのアテンションの変化に着目し，分析を行なった．24 層で構成される BrainBERT アーキテクチャでは，アテンションは 16 個のマルチヘッドを持っている． 本研究では，検証セットの下に 23 人の各被験者から 10 セットのデータを無作為に抽出して合計 230 セットとし，ユーザー別，アテンションの伝達，ブロック間のアテンションの変動を別々に分析した．ここで，python の可視化ツール bertviz パッケージ[1]を使用し，ヒートマップ生成方法をアテンションに合わせてカスタマイズしている． これにより，アテンションを定量的に抽出し，対応するグラフを生成して，結果がよりよく分析可能となった．

2.5 実験 4:BrainBERT によるテキストからの脳状態推定実験

BrainBERT の性能をさらに検証するために，テキストを使って脳状態を推定する実験も行う．単語の埋込ベクトルには BrainBERT と他の 20 個言語モデルを使用し，文の意味を表す `CLS` トークンから脳状態をリッジ回帰を用いて推定した．交差検証によってリッジ回帰其々の最適な正則化項を 100,000 と 10,000 の 2 つに決め，ボクセル毎に推定した値において $p = 0.01$ の両側検定で棄却されたボクセルのみを評価対象として，テストデータとピアソンの積率相関係数を用いて評価を行った．脳状態推定実験の結果によって，BERT を用いた脳状態推定の平均相関係数は 0.233%に対し，BrainBERT を用いた場合は 2.350%となり，相関が約 10 倍向上した．

3 おわりに

本研究では、言語刺激を受けたヒト脳の状態との対応関係を学習することで、脳活動データと BERT によって表現される言語の意味との共通の表現空間を捉える BrainBERT の構築を行なった．また，4 つの実験を通じて，BrainBERT の構築と有効性の検証を行なった．

今後，脳活動データを増やし BrainBERT の性能を向上させ，脳活動と言語による意味表現の間の定量関係を反映した BrainBERT の構築を目指したいと考えている．

参考文献

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.

[2] Mark A. Kramer. Nonlinear principal component analysis using autoassociative neural networks. *AIChE Journal*, Vol. 37, No. 2, pp. 233–243, 1991.

[3] Shohini Bhattasali, Jonathan Brennan, Wen-Ming Luh, Berta Franzluebbers, and John Hale. The alice datasets: fMRI & EEG observations of natural language comprehension. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pp. 120–125, Marseille, France, May 2020. European Language Resources Association.

[4] Francisco Pereira, Bin Lou, Brianna Pritchett, Samuel Ritter, Samuel Gershman, Nancy Kanwisher, Matthew Botvinick, and Evelina Fedorenko. Toward a universal decoder of linguistic meaning from brain activation. *Nature Communications*, Vol. 9, , 03 2018.

[1] https://github.com/jessevig/bertviz