脳内状態予測に特化した言語モデル構築への取り組み

濵田愛 [1]　田屋侑希 [1]　小林一郎 [1]
[1] お茶の水女子大学
{g2220532,g1620525,koba}@is.ocha.ac.jp

概要

本研究では，言語刺激と fMRI によって計測された脳活動データの対応関係を学習した言語モデル を構築し，得られた特徴量を用いて符号化モデルで脳内状態の予測を行う．とくに本研究では，脳活動データを言語モデルに入力する際の手法が脳内状態予測精度に与える影響を検証した．モデルの学習の際に入力する脳状態の抽出方法を変えて実験した結果，ピアソンの積率相関係数のボクセル平均は，事前学習済み BERT の特徴量を用いるベースラインが 0.197 と最も高かった．このことから，脳活動由来の情報を下流のリッジ回帰において活用可能な形式として獲得するための学習手法や，実験設定の最適化に課題が残されていることが示唆された．

1　はじめに

自然言語処理課題における脳活動と線形に対応することが報告されている [1]．言語刺激下における脳活動を予測する符号化モデル [2] は，刺激から脳活動を予測し，脳内表現に関する仮説の検証を可能にする．その実装として，言語モデルから得た特徴量を用いて，リッジ回帰などの線形モデルで各ボクセルを予測する枠組みが主流である．この枠組みは，学習の安定性や解釈可能性が高く，広く用いられている一方，そこで用いられる特徴表現は，あくまでも大量のテキストデータの統計的性質を学習した結果であり，脳活動の予測を目的として最適化されたものではない．そのため，言語処理タスク向けに固定された特徴量をそのまま用いる従来のアプローチでは，線形モデルによる予測精度の向上を阻害する要因となり得る [3]．そこで，本研究では，言語刺激の特徴量と脳活動データの対応関係を学習させたモデルの内部表現を利用することが，言語刺激に対する脳活動予測に影響を与えるかを検証する．特に，脳活動情報をモデルに組み込むことが予測精度および特徴表現の特性に与える影響を評価し，その要因について分析を行う．

2　関連研究

言語刺激下における脳活動を対象とした符号化モデルの研究では，言語刺激を入力として言語モデルの潜在層を通じて抽出した特徴量を説明変数とし，fMRI で観測した大脳皮質の各ボクセル応答を線形回帰 (主にリッジ回帰) で予測する符号化の枠組みが広く用いられてきた．これらの研究においては，言語モデルの層構造と脳内の処理階層との対応関係が示唆されており [1] [4]，使用する層や文脈長の選択が符号化性能に影響を与えることが報告されている [5]．この観点から，ボクセルごとに最適な層や文脈窓を選択することで予測精度を改善する手法が提案されている [6] が，改善は特徴量の「取り出し方」の範囲に限定される．一方で，脳活動データを用いて言語モデル自体を ファインチューニングし，内部表現を脳に整合させる試みも報告されている [7]．しかし一般に，大規模言語モデルに対する全パラメータ更新は，学習計算量に加えてタスクごとにモデル実体を保持する必要があり，計算資源およびメモリ／ストレージの観点から負担が大きいことが指摘されている [8, 9]．さらに，fMRI によって計測された脳活動データの取得は時間・費用コストが高く利用可能なデータ量が限られやすい点が指摘されている [10]．したがって，小規模データセットでの検証が中心となるが，その場合には大規模モデルが過学習しやすいことも報告されている [11]．

また，脳活動側の表現学習に注目した研究として，大規模な脳活動データを事前学習する基盤モデル [12] や，脳活動データを離散化したトークン系列として扱い，系列モデリングや言語との整合を図る仕組み [13] も提案されている．これらは汎用的な脳活動表現の獲得を目的とする点で本研究と関連するが，脳活動理解全般を対象とするものであり，本研究のように言語刺激下での脳内状態の予測精度の改善を直接の目的とはしていない．

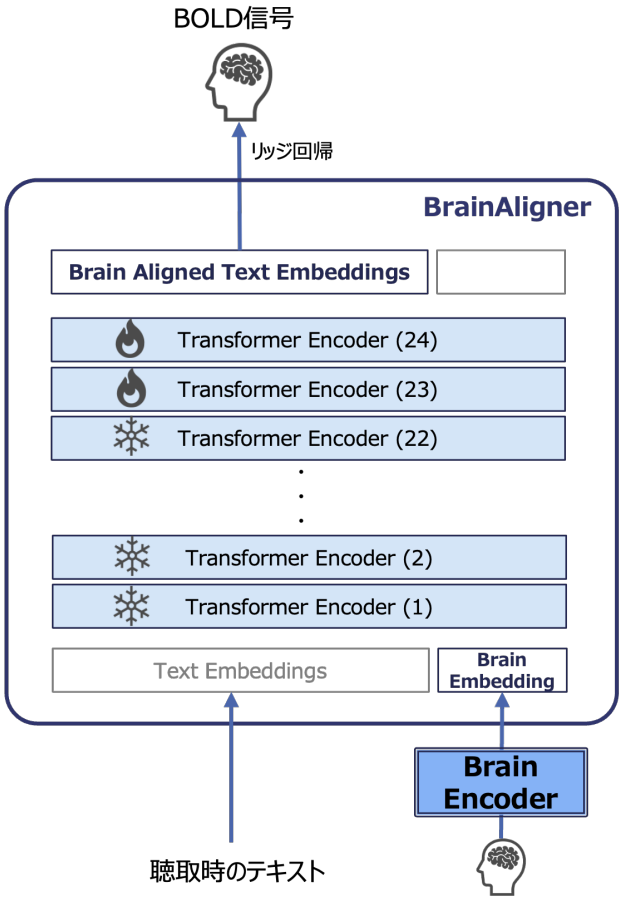

図 1　言語刺激と脳活動の対応関係を学習する言語モデルおよびそれを用いた脳内状態予測の概要図．学習時のみ Brain Encoder を活用して脳状態を入力し，評価時は聴取時のテキストのみ入力．

以上の知見を踏まえ，本研究では，言語刺激とその刺激下での脳活動データとの対応関係を事前に学習した言語モデルを構築し，そのモデルから抽出される特徴量を用いて符号化モデルによる脳内状態予測を行うことで，予測精度に対する有効性を検証する．さらに，学習時に与える脳活動データの設計が，脳内状態予測精度に及ぼす影響を分析する．

3　提案手法

本研究で構築する，言語刺激と脳活動の対応関係を学習する言語モデルおよびそれを用いた脳内状態予測の概要を図 1 に示す．学習時は，Brain Encoder（3.1 節参照）によって脳活動データの次元を削減し，圧縮した脳表現を BrainAligner（3.2 節参照）へ聴取時のテキストとともに入力し，これらの対応関係をとる表現学習を行う．評価時は，聴取時のテキストのみを入力とし，出力されたテキスト埋め込みを用いてボクセルごとのリッジ回帰で脳活動 (BOLD 信号) を予測し，予測値と実際の BOLD 信号の相関係数で評価する．本研究では，概要図の Brain Encoder に着目し，高次元の脳活動データを圧縮する際に，異なる 3 つの Brain Encoder を用いることで，符号化モデルによる BOLD 信号の予測精度に与える影響を比較する．

3.1　Brain Encoder：脳状態の次元削減

大脳皮質の脳活動データは数万次元のボクセル空間で与えられるため，言語モデルである BrainAligner の入力として扱うには次元数が大きく，圧縮が必要となる．本研究では，以下の 3 つの手法によって次元数を減らし，それぞれの性能を比較する．

3.1.1　脳活動事前学習モデルの内部表現抽出

自己教師あり学習によって大規模脳活動データから汎用的な表現を獲得するモデル BrainLM [12] がある．本研究では，BrainLM を脳活動の特徴抽出モデルとして用い，使用する脳活動の時系列に対して，各時間点 (TR) に対応する内部表現を取得する．すなわち，時刻 t における BrainLM の隠れ状態を $\mathbf{h}_t \in \mathbb{R}^{1280}$ とし，この $\mathbf{h}_t$ を脳状態表現として利用する．ただし，BrainLM から得られる 1280 次元の内部表現に対して線形写像による次元圧縮を行い，1024 次元にして用いる．具体的には，学習可能な重み行列 $\mathbf{W} \in \mathbb{R}^{1024\times1280}$ とバイアス $\mathbf{b} \in \mathbb{R}^{1024}$ を導入し，

$$\mathbf{z}_t = \mathbf{W}\mathbf{h}_t + \mathbf{b} \in \mathbb{R}^{1024} \tag{1}$$

として 1024 次元表現 $\mathbf{z}_t$ を得る．以降の解析ではこの $\mathbf{z}_t$ を用いる．なお，本研究の目的は BrainLM 自体の再学習ではなく，事前学習済みモデルが保持する表現を下流に転用する点にあるため，BrainLM 本体は固定し，線形変換層 (式 1) のみを下流設定に合わせて学習する設計とする．本研究ではこの圧縮方法を「BrainLM」と表記する．

3.1.2　Explainable Variance に基づくボクセル選択

今回使用するデータセットに含まれる反復計測に基づき，観測変動のうち信号として再現可能な成分が占める割合である Explainable Variance (EV) [14] を算出し，EV の高いボクセル上位 1 万個を選択する．選択されたボクセルの時系列を Z-score を用いて正規化，線形層で 1024 次元に線形変換し BrainAligner への脳活動データの入力とする．これによりノイズの大きいボクセルを抑制し，下流の学習・評価の安

定性を高めることを可能とする．本研究ではこの圧縮方法を「High_EV」と表記する．

3.1.3 言語刺激で説明可能なボクセルの選択

この手法において，まず本研究とは別に，事前学習済み Bert-large-uncased の埋め込みベクトルを言語刺激の特徴量とし，大脳皮質の各ボクセルを予測する線形リッジ回帰による符号化モデルを構築する．その符号化モデルによる予測において，相関が高いボクセルの上位 1 万個をブロック置換検定と FDR 補正により有意判定して選択する．これを下流で線形層により 1024 次元へ線形変換する．これにより，言語刺激への反応が有意なボクセルに基づいた予測が可能になる．本研究ではこの圧縮方法を「ES（Encoding Selection）」と表記する．

3.2 BrainAligner

BrainAligner は，事前学習済み BERT-large-uncased を初期重みとし，3.1 節の 3 種類の方法で次元圧縮した脳活動データを入力した上で Masked Language Modeling (MLM) を行いファインチューニングする．自然言語聴取 fMRI データセット [15](4.1 節にて説明) におけるテキスト刺激と対応する脳活動データを用い，下位層を凍結し，上位 2 層のみを学習することで，一般言語能力の保持と過学習抑制，ならびに計算コスト低減を狙う．最適化手法には AdamW ($\beta = (0.9, 0.98)$，重み減衰を 0.01) を採用し，最大系列長は 256 とし，学習率を 1×10^{-5}(ウォームアップ 2,000 ステップ) に設定した．正則化としてドロップアウト率を 0.1 とした．また，ミニバッチサイズは 32 とし，勾配蓄積 2 によりパラメータ更新を行った (更新あたりの実効バッチサイズは 32×2)．MLM 学習では BERT [16] のデフォルト設定に従い，入力系列中のトークナイザによって分割されたトークンのうち 15% を予測対象としてランダムに選択した．選択したトークンについては，80% の確率でマスクトークン (`[MASK]`) に置換し，10% の確率で語彙からサンプリングした別トークンに置換し，残り 10% は元のトークンのまま保持した．損失は，選択されたトークン位置に対してのみ計算した．学習時は脳活動データ を入力することによりテキストと脳活動データの対応関係がとれた新たな埋め込みベクトルが獲得される．評価時は，脳活動データの取得単位である 1TR に対応するテキストのみを入力とし，出力された埋め込みベクトルを平均プーリングし，BrainAligner の出力とする．

3.3 下流評価：ボクセルごとのリッジ回帰

得られた 1TR 毎のテキスト埋め込み x を説明変数，各ボクセルの BOLD 信号 y を目的変数として，リッジ回帰によりボクセルごとの符号化モデルを学習する．評価はテストデータでピアソンの積率相関係数 r をボクセルごとに算出し，その平均値・中央値を計算する．正則化係数 α は検証用データで選択し，全ボクセル共通の α を用いる設定とする (候補集合は $\{0.01, 0.1, 1, 10, 100, 1000\}$) .

4 実験

4.1 使用データセット

本研究では，Huth のグループによる自然言語聴取課題の fMRI 脳活動データ [15] を用い，被験者 3 名 (UTS01–03) が計 84 本の物語を聴取している際に計測されたものを使用する．聴取時のテキストと脳活動データの時間対応は，データセットに含まれる TextGrid を用いて構成する．TextGrid とは，音声刺激の書き起こしに対して，各単語の開始・終了時刻（単語境界）を付与したアノテーションである．具体的には，各単語の時間区間と脳活動データの取得時刻の対応関係から，TR 単位のテキスト系列を作成する．標準的な前処理 (モーション補正，空間正規化，スライスタイミング補正など) があらかじめ適用されているデータを用いるため．本研究では，各ボクセルごとの Z-score 正規化のみ追加で適用した．データ中に含まれる NaN/Inf などの非有限値は，解析前に 0 へ置換する処理を施した．なお，実験で用いた脳活動データは言語理解に関与する主要領域である皮質部分のみを用いた．

リッジ回帰の学習にあたり，84 本の物語データを学習・検証・テスト用にそれぞれ 60 本，8 本，16 本に分割する．さらに学習用の 60 本のデータを 56：4 で分割し，BrainAligner の学習および検証に用いる．

4.2 比較手法

提案手法の BrainAligner による性能向上を評価するため，脳活動データに基づくファインチューニングを行わずに，汎用言語モデルが出力するテキスト埋め込みをそのまま用いる方法をベースライン (**Baseline**) とする．具体的には，事前学習済み BERT-large-uncased から各 TR に対応するテキスト

表 1　各 Brain Encoder で圧縮した脳活動データを用いて学習した BrainAligner から得た特徴量を元に，符号化モデルで脳内状態予測をし，実測値との相関係数を計算した結果．相関係数の平均値と中央値について，被験者ごとの結果と被験者 3 名の平均を示す．

手法	被験者	r_mean	r_median
Baseline	UTS01	0.196112	0.195064
	UTS02	0.196028	0.194396
	UTS03	**0.199710**	**0.196044**
	平均	**0.197284**	**0.195168**
BrainLM	UTS01	0.193483	0.193185
	UTS02	0.193278	0.192930
	UTS03	0.195252	0.194018
	平均	0.194004	0.193378
High_EV	UTS01	0.170763	0.170271
	UTS02	0.169844	0.168978
	UTS03	0.172562	0.170609
	平均	0.171056	0.169953
ES	UTS01	0.171014	0.170423
	UTS02	0.170430	0.169498
	UTS03	0.172475	0.170414
	平均	0.171306	0.170112

埋め込みを抽出し，その埋め込みから各ボクセルの BOLD 信号を予測するリッジ回帰モデルを学習する．提案手法である BrainAligner については，Brain Encoder の 3 つの手法である BrainLM（3.1.1 節），High_EV（3.1.2 節）, ES（3.1.3 節）を比較する．テストデータを用いて，各ボクセルについて時系列データである真の BOLD 信号と予測した BOLD 信号の相関係数を算出し，その平均・中央値を求める．

4.3　実験結果

本研究では，言語刺激に対する脳内状態予測（符号化）性能を，各ボクセルにおけるピアソンの積率相関係数 r により評価した．表 1 に，各手法において，全ボクセルの相関係数 r の平均値（r_{mean}）および中央値（r_{median}）について，被験者別（UTS01–03）の結果と，被験者 3 名の結果を平均したものを併せて示す．平均の結果では Baseline が最も高い相関を示し（$r_{\text{mean}} = 0.1973$，$r_{\text{median}} = 0.1951$），次いで BrainLM が近い値を示した（$r_{\text{mean}} = 0.1940$，$r_{\text{median}} = 0.1933$）．一方，ES および High_EV は相関が低く（それぞれ $r_{\text{mean}} = 0.1713$，0.1711），Baseline・BrainLM との差が確認された．Baseline と BrainLM の差は小さく，3 名平均における差分は r_{mean} で約 0.0033，r_{median} で約 0.0018 にとどまった．被験者別に見ても手法間の相対的な傾向は概ね一貫しており，いずれの被験者においても Baseline が最も高い相関を示し，BrainLM がそれに続いた．

5　考察

提案手法と Baseline の間に有意な性能差が見られなかった要因として，脳活動データの導入効果が，下流のリッジ回帰において抽出可能な形として表現空間へ十分に反映されなかった可能性が考えられる．本研究では，BrainAligner の学習に MLM を採用したが，脳活動由来の情報を下流のリッジ回帰でも有効に活用できる特徴表現を得るためには，この枠組みだけでは不十分であった可能性がある．脳活動データとテキストの対応関係をより直接的に学習するためには，既存の MLM の適用にとどまらず学習手法の改善や新たな目的関数の導入が必要であると考えられる．また，被験者間で手法の相対的な順位が概ね一致した点は，結果に対する支配的な要因が被験者固有の個人差にあるのではなく，各比較手法にあることを示唆する．以上より，本研究では言語モデルへの脳活動データ導入による符号化性能の一貫した向上は確認されなかったものの，これらの結果は提案手法の枠組み自体の有効性を否定するものではなく，前述した学習手法の改善や条件設定の改善による発展の余地を残すものであるといえる．

6　おわりに

本研究では，自然言語聴取 fMRI 脳活動データセットを対象に，脳活動データを導入して学習を行ったモデル（BrainAligner）を提案した．本研究の狙いは，下流の符号化モデルの工夫に依存して予測性能を高めるのではなく，言語表現学習（MLM）の段階で脳活動情報を組み込み，言語刺激に対する脳内状態予測に有用な言語特徴量を獲得する点にあった．実験の結果，BrainLM で圧縮した脳活動データを用いた学習により得られた言語特徴量は，Baseline と比較して性能の著しい低下は示さなかったものの，符号化性能を明確に改善するには至らなかった．考察で述べたように，これらは脳活動由来の情報を下流のリッジ回帰においても有効に活用可能な形式として獲得するための学習手法や，実験設定の最適化に課題が残されていることを示唆する．今後は，実験設定の改善に加え，関心領域（ROI）情報の活用や MLM の枠組みを超えた目的関数の導入に取り組み，脳と言語モデルの整合性を高めるための手法を発展させていきたい．

参考文献

[1] Charlotte Caucheteux and Jean-Rémi King. Brains and algorithms partially converge in natural language processing. *Communications biology*, Vol. 5, No. 1, p. 134, 2022.

[2] Thomas Naselaris, Kendrick N. Kay, Shinji Nishimoto, and Jack L. Gallant. Encoding and decoding in fMRI. *NeuroImage*, Vol. 56, No. 2, pp. 400–410, 2011.

[3] Richard Antonello, Aditya Vaidya, and Alexander Huth. Scaling laws for language encoding models in fmri. *Advances in Neural Information Processing Systems*, Vol. 36, pp. 21895–21907, 2023.

[4] Sreejan Kumar, Theodore R Sumers, Takateru Yamakoshi, Ariel Goldstein, Uri Hasson, Kenneth A Norman, Thomas L Griffiths, Robert D Hawkins, and Samuel A Nastase. Shared functional specialization in transformer-based language models and the human brain. *Nature communications*, Vol. 15, No. 1, p. 5523, 2024.

[5] Shailee Jain and Alexander Huth. Incorporating context into language encoding models for fmri. *Advances in neural information processing systems*, Vol. 31, , 2018.

[6] Anuja Negi, Christine Tseng, Anwar O. Nunez-Elizalde, Xue Lily Gong, and Fatma Deniz. Optimizing language model embeddings to voxel activity improves brain activity predictions. *bioRxiv*, 2025.

[7] Anuja Negi, Subba Reddy Oota, Anwar O Nunez-Elizalde, Manish Gupta, and Fatma Deniz. Brain-informed fine-tuning for improved multilingual understanding in language models. *bioRxiv*, pp. 2025–07, 2025.

[8] Ning Ding, Yujia Qin, Guang Yang, Fuchao Wei, Zonghan Yang, Yusheng Su, Shengding Hu, Yulin Chen, Chi-Min Chan, Weize Chen, et al. Parameter-efficient fine-tuning of large-scale pre-trained language models. *Nature Machine Intelligence*, Vol. 5, pp. 220–235, 2023.

[9] Edward Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022.

[10] Navve Wasserman, Roman Beliy, Roy Urbach, and Michal Irani. Functional brain-to-brain transformation with no shared data. *arXiv preprint arXiv:2404.11143*, 2024.

[11] Lev Kiar Avberšek and Grega Repovš. Deep learning in neuroimaging data analysis: Applications, challenges, and solutions. *Frontiers in Neuroimaging*, 2022.

[12] Josue Ortega Caro, Antonio Henrique de Oliveira Fonseca, Syed Asad Rizvi, Matteo Rosati, Christopher L. Averill, James L. Cross, Prateek Mittal, Emanuele Zappala, Rahul Madhav Dhodapkar, Chadi Abdallah, and David van Dijk. BrainLM: A foundation model for brain activity recordings. In *International Conference on Learning Representations (ICLR)*, 2024.

[13] Yuxiang Wei, Yanteng Zhang, Xi Xiao, Chengxuan Qian, Tianyang Wang, and Vince D Calhoun. fmri-lm: Towards a universal foundation model for language-aligned fmri understanding. *arXiv preprint arXiv:2511.21760*, 2025.

[14] Alexander G. Huth, Shinji Nishimoto, An T. Vu, and Jack L. Gallant. A continuous semantic space describes the representation of thousands of object and action categories across the human brain. *Neuron*, Vol. 76, No. 6, pp. 1210–1224, 2012.

[15] Amanda LeBel, Lauren Wagner, Shailee Jain, Aneesh Adhikari-Desai, Bhavin Gupta, Allyson Morgenthal, Jerry Tang, Lixiang Xu, and Alexander G. Huth. A natural language fMRI dataset for voxelwise encoding models. *Scientific Data*, Vol. 10, No. 1, p. 555, 2023.

[16] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT 2019*, pp. 4171–4186, 2019.