脳アトラス・脳基盤モデルを用いた脳-テキストデコーディング

赤間美香　吉田遼　大関洋平
東京大学
{haruka-akama,yoshiryo0617,oseki}@g.ecc.u-tokyo.ac.jp

概要

言語を知覚したときの脳データから、自然言語処理の技術を用いて、テキストを復元する技術を脳-テキストデコーディングという。脳データから良質な脳表現を獲得することは、脳-テキストデコーディングの性能に直結する重要な課題である。本研究では、良質な脳表現を獲得するために、「脳アトラスによる次元削減」と「脳基盤モデル」を導入することを提案し、それぞれの手法の有効性を検証した。実験の結果、前者は有効であったが、後者による性能向上は観察されなかった。この結果は、モデルの初期段階における次元削減においてなお大きな改善余地が残されていることや言語タスクに特化した大規模脳表現モデルの構築の必要性を示唆する。

1　はじめに

私たちは通常、発話や手話といった運動を介して自然言語を表現する。しかし、これらの出力手段に依存せず、脳内に表象された言語情報を直接復号することも可能であろうか。このような脳内言語表現の言語処理に取り組む研究分野・技術を、脳-テキストデコーディングという。

被験者の身体を傷つけることなく記録された非侵襲脳データを用いた脳-テキストデコーディングは、長らく単語レベルで発展してきた [1]。近年、デコーダとして大規模言語モデル (Large Language Model, LLM) を援用することで、文章レベルでのデコーディングも可能になりつつある [2]。しかしながら、デコーダが高性能であっても、エンコーダが脳データから良質な脳表現を獲得できなければ、性能向上のボトルネックとなり得ることから、エンコーダの性能向上も急務である。

そこで本研究では、(1) 脳アトラスによる次元削減、(2) 脳基盤モデルの導入により、良質な脳情報を獲得できるという仮説を立て、それぞれの手法の有効性を検証する。先行研究 [2] の脳表現を抽出するための次元削減手法およびエンコーダモデルの設計に注目すると、高次元の脳データに対して主成分分析（PCA）が用いられているが、この方法では脳の解剖学的・空間的構造情報が失われる可能性がある。また、エンコーダには単純な多層パーセプトロン（MLP）のみが用いられており、複雑な脳情報を十分に活用できていない可能性がある。

2　関連研究

非侵襲脳データから「文章レベル」で言語情報を復元することは長らく困難な課題とされてきた。その要因の１つとして、非侵襲計測における、どれくらいの時間的な精度で測定できるかという時間解像度とどれくらいの空間的精度で記録できるかという空間解像度のトレードオフが挙げられる。また、脳活動とテキストというモダリティの隔たりも大きな障害である。しかし近年、機能的磁気共鳴画像法（functional Magnetic Resonance Imaging, fMRI）による非侵襲脳データに対し、デコーダとして LLM を導入することで、被験者が聞いた物語を文章レベルで生成的に復元できることが示された [2]。この手法では、LLM が大規模コーパスから獲得した言語の事前分布を活用し、脳データから捕捉できなかった情報を補完することで、整合的な文章生成を可能にしている。

同研究で提案された手法では、ターゲット文に対応する fMRI データと、その直前部分のテキストプレフィックスを入力とする。fMRI データは主成分分析（PCA）が適用され次元削減される一方、テキ

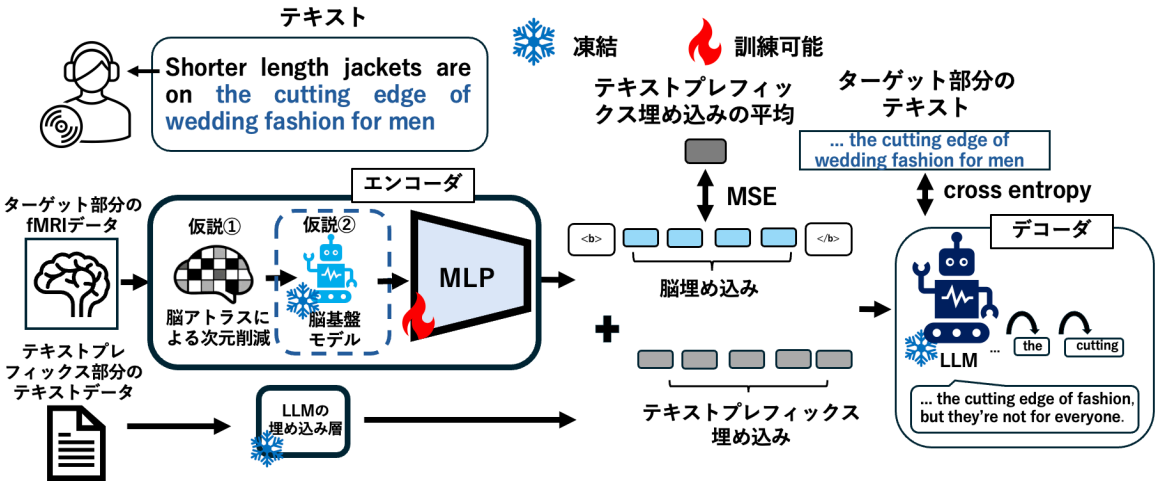

図 1: 本研究の概要

ストは LLM の埋め込み層によって埋め込み表現に変換される。そして、次元削減後の脳データは、多層パーセプトロン（MLP）によって LLM のテキスト埋め込み空間に写像される。この脳埋め込みとテキスト埋め込みは、特殊トークンを付加したうえで横方向に連結され、LLM の入力として与えられる。

文章理解に関わる脳活動は、時間的に連続した文脈依存的処理であり、複数の皮質領域にまたがる分散表現として実現される [3]。このように脳表現は複雑な構造を内包しているため、低次元表現への変換過程でこれらの情報が失われると、復元性能のボトルネックとなりうる。先行研究で用いられる PCA は分散最大化に基づく線形次元削減手法であり、脳の解剖学的配置や局所構造を明示的には考慮しない。また、PCA により得られた低次元表現を LLM 入力空間へ写像する小規模 MLP は、次元整合の役割は果たすものの、脳データの時間的文脈や脳の領域間の脳活動の関係性を十分に表現できない可能性がある。以上より、文章レベルの脳-テキストデコーディングにおける性能向上の鍵は、デコーダの改良のみならず、脳表現をいかに構成するかというエンコーディング段階にあると考えられる。

3 脳情報エンコーディング

前節で述べた課題意識に基づき、本研究では、(1) 解剖学的に妥当な次元削減を行うこと、および (2) より大規模かつ表現力の高いエンコーダモデルを導入することによって、脳情報をより効果的に活用できるという二つの仮説を検証する。具体的には、仮説 (1) の検証のために AAL-424 アトラス [4] を、仮説 (2) の検証のために脳基盤モデル [5] を導入する。

3.1 脳アトラス

脳アトラスとは、脳構造や脳機能のさまざまな側面、およびそれらの関係性を記述した脳の地図である [6]。本論では脳の解剖学的側面を切り取って構成された解剖学的アトラスの一つである AAL-424 アトラスを採用した。AAL（Automated Anatomical Labeling）とは、複数被験者の脳の構造を捉えた磁気共鳴画像 (Magnetic Resonance Imaging, MRI) を平均化した脳テンプレートに基づき、大脳皮質および皮質下構造を解剖学的に分割し、脳データの最小単位である各ボクセルに一意な解剖学ラベルを自動的に割り当てるため脳アトラスである [7]。AAL-424 とは、424 人分の脳テンプレートに基づく解剖学的アトラスを指す。検証では、このアトラスに基づいて fMRI データに含まれる全ボクセルをアトラス空間へ再配置し、各領域ごとに集約することで、ボクセル次元からアトラス領域数への次元削減を行う。

3.2 脳基盤モデル

脳基盤モデルとは、多様な下流タスクに応用可能な汎用的脳表現の学習を目的として、約 6,700 時間に及ぶ大規模 fMRI データ [8] を用いて学習された、大規模な脳表現モデルである。学習フレームワークは BERT [9] や Vision Transformer [10, 11] に着想を得

たマスクモデリングに基づいている。これにより、脳内の複数の領域にまたがる空間的な広がりと、時間とともに変化する活動パターンが重なり合って構成される、脳信号特有の複雑な空間的・時間的ダイナミクスを自己教師あり学習によって捉えることが可能となっている。推論時には、AAL-424 アトラスで次元削減された 424 次元の脳データを入力として受け取り、その時空間的文脈に依存した脳信号の再構築結果を出力する。本研究では、この脳基盤モデルを事前学習済みの状態で凍結し、脳基盤モデルが再構築した脳信号をデコーダに渡す。これにより、エンコーダのパラメータの拡大と、時空間構造を考慮した再構築によるノイズ低減が期待される。

4 実験設定

実験では、(A) エンコーダが PCA による次元削減と MLP であるベースラインと、(B) エンコーダが解剖学的に妥当な次元削減と MLP である設定と、(C) エンコーダが解剖学的に妥当な次元削減と脳基盤モデルと MLP である設定の 3 つの評価指標に基づき、それぞれのモデルの復元精度によって比較する。

モデル MLP は隠れ層 3 層からなる構成とし、各層の活性化関数には ReLU6 を用いた。LLM は、全ての設定において、GPT2 シリーズ[1)]の 124M の GPT2-small を使った。設定 (C) でのみ約 19M の凍結した脳基盤モデルを用いた。

学習データ 約 4.6 時間分の物語を聞いているときの fMRI データセットである Narratives データセット [12] を先行研究 [2] に基づき 27 人分抽出して用いた。このデータセットはモーション補正や正則化などの標準的な前処理済みの公開データセットである。

学習 脳基盤モデルおよび LLM のパラメータはすべて凍結し、MLP と特殊トークンのみを学習対象とした。学習は先行研究 [2] に従い、2 段階で行った。第 1 段階では、脳表現とテキスト埋め込み空間との整合を目的とした。脳基盤モデルから得られる脳表現を $\mathbf{B} \in \mathbb{R}^{T\times 424}$ とし、これを MLP によって LLM の埋め込み次元 d に射影したベクトルを $\mathbf{z}_{\text{brain}} \in \mathbb{R}^d$ とする。一方対応するテキストプレフィックス埋め込みを $\{\mathbf{e}_1, \ldots, \mathbf{e}_N\}$ としその平均ベクトルを $\mathbf{z}_{\text{prefix}} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{e}_i$ と定義する。このとき MLP の出力がテキスト埋め込み空間に整合するよう以下の平均二乗誤差 (Mean Squared Error; MSE) を最小化した：

$$\mathcal{L}_{\text{MSE}} = \left\|\mathbf{z}_{\text{brain}} - \mathbf{z}_{\text{prefix}}\right\|_2^2$$

この段階では異なるモダリティ間の表現空間を整合させるための初期化のみを行った。第 2 段階では、第 1 段階で整合された 2 つのモダリティの埋め込みを連結した入力列を条件として、LLM による文章生成を行った。具体的には、LLM への入力埋め込み列 $\mathbf{X}$ を $\mathbf{X} = \left[\langle\text{brain}\rangle, \mathbf{z}_{\text{brain}}, \langle/\text{brain}\rangle, \mathbf{e}_1, \ldots, \mathbf{e}_N\right]$ と定義する。ここで $\langle\text{brain}\rangle$ および $\langle/\text{brain}\rangle$ は脳埋め込みの境界を示す特殊トークンである。

生成対象となる正解トークン列を $\mathbf{y} = (y_1, \ldots, y_M)$ とすると、モデルは各時刻 t において、入力列 $\mathbf{X}$ とそれまでに生成されたトークン $y_{<t}$ を条件として、次トークンの条件付き確率 $p(y_t \mid y_{<t}, \mathbf{X})$ を推定する。このとき、正解文に対する尤度を最大化するため、以下のクロスエントロピー損失を最小化した：

$$\mathcal{L}_{\text{CE}} = -\sum_{t=1}^{M} \log p(y_t \mid y_{<t}, \mathbf{X})$$

本研究における学習設定は以下のとおりである。ミニバッチサイズは 1 とし、エポック数は 100 に設定した。最適化に用いる学習率は 1.03×10^{-5} とし、L2 正則化係数は 1.16×10^{-3} とした。また、過学習を抑制するためドロップアウト率を 0.57 に設定した。なお、検証性能の改善が 10 エポック連続して観測されない場合には、学習を早期終了する設定とした。ハイパーパラメータ探索には Optuna [13] を用いた。

評価手法 評価には主に翻訳性能を評価する次の評価指標を用いた。

- **BLEU (Bilingual Evaluation Understudy)** [14]：生成された文章と参照文の一致度を「適合率 (Precision)」に基づいて評価し、短すぎる文章に対する罰則を加味して算出される指標である。
- **ROUGE (Recall-Oriented Understudy for Gisting Evaluation)** [15]：ROUGE は参照文に含まれる

1) https://huggingface.co/openai-community/gpt2

語句が生成文にどれだけ含まれているかという「再現率（Recall）」に焦点を当て、内容の網羅性を評価している。特に、ROUGE-1 は共通して出現する単一単語の個数に基づき、単語レベルでの内容の重なりを評価する指標である。また、ROUGE-L は共通する「最長共通部分列」の長さにより、単語の出現順序を考慮した文構造の類似性を評価する指標である。

- **WER (Word Error Rate)** [16]：参照文を正解とした際に、生成文を一致させるために必要な置換・削除・挿入の操作回数に基づき、単語レベルでの誤り率を測定する指標である。

5　結果・考察

5.1　仮説①：脳アトラスによる次元削減

Metric	(A) ベースライン	(B)424-atlas	(C) 脳基盤モデル
BLEU-1	11.35	12.84*	12.59
ROUGE-1	10.84	12.09*	11.77
ROUGE-L	10.15	11.38*	11.10
WER（↓）	95.38	94.37*	94.49
Validation Loss	5.42	5.24*	5.13†

表 1: （A）エンコーダが PCA による次元削減と MLP であるベースライン設定と（B）エンコーダを解剖学的に妥当な次元削減と MLP にした設定と（C）エンコーダを解剖学的に妥当な次元削減と脳基盤モデルと MLP にした設定の性能比較。
統計検定には Wilcoxon の符号付順位検定（両側検定）を用いた。各指標について、（A）と（B）の対応する試行ペア間と（B）と（C）の対応する試行ペア間とで検定を行い、ゼロ差分は Pratt 法 [17] で処理し連続性補正を適用した。複数指標に対する多重比較の影響を抑えるため、Benjamini–Hochberg 法 [18] による FDR 補正を行い有意水準は $q < 0.05$ とした。（B）の (A) に対する優位差は＊で示し、（C）の（B）に対する優位差は†で示した。

実験の結果（表 1）、AAL-424 アトラスに基づいて解剖学的に妥当な次元削減を行うと、単に PCA で次元削減するベースラインよりも、全ての言語評価指標において統計的に有意な性能向上が見られた。これは脳の本来持つ空間情報を保存したまま次元削減をすることで、それらの情報を活かした脳表現出力が得られるからだと考えられる。言語理解や意味処理に関わる脳活動は単一の局所的領域に局在するのではなく、複数の皮質領域にまたがる空間的に分散した活動パターンとして表現されることが示されている [19, 20]。このことから、脳活動の空間構造を保持した表現は、脳-テキストデコーディングにおいて重要な役割を果たすと考えられる。また、この結果は、モデルの初期段階における次元削減においてなお大きな改善余地が残されている事を示唆している。

5.2　仮説②：脳基盤モデル

表 1 より、脳アトラスで 424 次元に次元削減した脳データを脳基盤モデルに入力しても、すべての言語評価指標において性能向上は確認されなかった。この結果は、単にエンコーダの規模を拡大し、再構築された脳信号を用いるだけでは、下流の言語デコーディング性能の改善には直結しないことを示している。言語タスクにおいて重要なのは、刺激や課題に応じて変化する脳活動の空間的分布構造と、その時間的推移をモデルが適切に表現・統合し、タスク関連表現としてエンコーダに導入できているかどうかである。本研究で用いた脳基盤モデルは、安静時 fMRI や顔や図形から感情を推定する非言語的タスクのデータで学習されており、言語処理特有の時空間ダイナミクスを十分に獲得できなかった可能性が高い。一方で、検証損失は有意に減少しており、脳基盤モデルの導入が学習の安定性を向上させたことが示された。このことから、脳基盤モデルはタスク特異的情報の強化には至らないものの、脳活動に共通するある種の普遍的表現を使ってノイズ除去や正則化に近い役割を果たしていたと考えられる。以上から、言語タスクに特化した大規模脳表現モデルの必要性が示唆される。

6　おわりに

本研究では、脳-テキストデコーディングにおいて、脳アトラスによる次元削減と脳基盤モデルの導入が、良質な脳表現の獲得および性能向上に繋がるか検証した。実験の結果、脳アトラスによる次元削減は有効であった一方で、脳基盤モデルの能力は十分に活用できなかったことが示された。今後は、言語タスクに特化した大規模な脳表現モデルの構築や脳-テキストデコーディングに最も適したアトラスを探索を行いたい。

謝辞

本研究は，JSPS 科研費 JP24H00087，JST さきがけ JPMJPR21C2，JST CREST JPMJCR2565，JST BOOST JPMJBY24B2 の支援を受けたものです．

参考文献

[1] Francisco Pereira, Bin Lou, Brianna Pritchett, Samuel Ritter, Samuel J Gershman, Nancy Kanwisher, Matthew Botvinick, and Evelina Fedorenko. Toward a universal decoder of linguistic meaning from brain activation. **Nature communications**, Vol. 9, No. 1, p. 963, 2018.

[2] Ziyi Ye, Qingyao Ai, Yiqun Liu, Maarten de Rijke, Min Zhang, Christina Lioma, and Tuukka Ruotsalo. Generative language reconstruction from brain recordings. **Communications Biology**, Vol. 8, No. 1, p. 346, 2025.

[3] James B Heald, Daniel M Wolpert, and Máté Lengyel. The computational and neural bases of context-dependent learning. **Annual Review of Neuroscience**, Vol. 46, No. 1, pp. 233–258, 2023.

[4] Samaneh Nemati, Teddy J Akiki, Jeremy Roscoe, Yumeng Ju, Christopher L Averill, Samar Fouda, Arpan Dutta, Shane McKie, John H Krystal, JF William Deakin, et al. A unique brain connectome fingerprint predates and predicts response to antidepressants. **IScience**, Vol. 23, No. 1, 2020.

[5] Josue Ortega Caro, Antonio H de O Fonseca, Christopher Averill, Syed A Rizvi, Matteo Rosati, James L Cross, Prateek Mittal, Emanuele Zappala, Daniel Levine, Rahul M Dhodapkar, et al. Brainlm: A foundation model for brain activity recordings. **bioRxiv**, pp. 2023–09, 2023.

[6] Isaac Bankman. **Handbook of medical image processing and analysis**. Elsevier, 2008.

[7] Nathalie Tzourio-Mazoyer, Brigitte Landeau, Dimitri Papathanassiou, Fabrice Crivello, Octave Etard, Nicolas Delcroix, Bernard Mazoyer, and Marc Joliot. Automated anatomical labeling of activations in spm using a macroscopic anatomical parcellation of the mni mri single-subject brain. **Neuroimage**, Vol. 15, No. 1, pp. 273–289, 2002.

[8] Karla L Miller, Fidel Alfaro-Almagro, Neal K Bangerter, David L Thomas, Essa Yacoub, Junqian Xu, Andreas J Bartsch, Saad Jbabdi, Stamatios N Sotiropoulos, Jesper LR Andersson, et al. Multimodal population brain imaging in the uk biobank prospective epidemiological study. **Nature neuroscience**, Vol. 19, No. 11, pp. 1523–1536, 2016.

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

[10] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. **arXiv preprint arXiv:2010.11929**, 2020.

[11] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In **Proceedings of the IEEE/CVF conference on computer vision and pattern recognition**, pp. 16000–16009, 2022.

[12] Samuel A Nastase, Yun-Fei Liu, Hanna Hillman, Asieh Zadbood, Liat Hasenfratz, Neggin Keshavarzian, Janice Chen, Christopher J Honey, Yaara Yeshurun, Mor Regev, et al. The“narratives”fmri dataset for evaluating models of naturalistic language comprehension. **Scientific data**, Vol. 8, No. 1, p. 250, 2021.

[13] Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A next-generation hyperparameter optimization framework. In **Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining**, pp. 2623–2631, 2019.

[14] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In **Proceedings of the 40th annual meeting of the Association for Computational Linguistics**, pp. 311–318, 2002.

[15] Shweta Chauhan and Philemon Daniel. A comprehensive survey on various fully automatic machine translation evaluation metrics. **Neural Processing Letters**, Vol. 55, No. 9, pp. 12663–12717, 2023.

[16] Dietrich Klakow and Jochen Peters. Testing the correlation of word error rate and perplexity. **Speech Communication**, Vol. 38, No. 1-2, pp. 19–28, 2002.

[17] John W Pratt. Remarks on zeros and ties in the wilcoxon signed rank procedures. **Journal of the American Statistical Association**, Vol. 54, No. 287, pp. 655–667, 1959.

[18] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. **Journal of the Royal statistical society: series B (Methodological)**, Vol. 57, No. 1, pp. 289–300, 1995.

[19] James V Haxby, M Ida Gobbini, Maura L Furey, Alumit Ishai, Jennifer L Schouten, and Pietro Pietrini. Distributed and overlapping representations of faces and objects in ventral temporal cortex. **Science**, Vol. 293, No. 5539, pp. 2425–2430, 2001.

[20] Jerry Tang, Amanda LeBel, Shailee Jain, and Alexander G Huth. Semantic reconstruction of continuous language from non-invasive brain recordings. **Nature Neuroscience**, Vol. 26, No. 5, pp. 858–866, 2023.