

[研究助成 (C)]

能動的なヒト脳活動における脳内言語情報解読手法開発への挑戦

2257015

研究代表者
(助成金受領者)

共同研究者

お茶の水女子大学
大学院人間文化創成科学研究科

お茶の水女子大学

情報通信研究機構
未来 ICT 研究所
脳情報通信融合研究センター

博士後期課程

教授

特別研究員

LIU MUXUAN

小 林 一 郎

西 田 知 史

[研究の目的]

本研究の目的は、画像を観察した際に被験者が心の中で想起する言語内容を、非侵襲的に計測された fMRI 脳活動データから推定するための基盤的手法を構築することである。脳活動から意味内容や言語内容を解読する研究は近年進展しており、とくに音声や文章のような言語刺激を対象とした研究では、刺激系列が明確な時間軸を持つため、脳信号との対応を比較的取りやすい。実際、音声刺激データでは、BOLD 信号の遅延補正やスライディングウィンドウを用いて、単語単位の言語特徴と脳応答との対応を調べる研究が進められてきた。

しかし、画像観察時に被験者が心の中で自然に想起する文を対象とする場合、従来研究とは異なる難しさがある。この課題には音声刺激のような明確な時間軸がなく、fMRI 信号は固定長の時間系列として取得される一方、想起文の長さは一定ではない。そのため、固定長の脳活動系列と可変長の言語表現をどのように対応づけるかが大きな問題となる。また、対象は一般的な画像キャプションではなく、被験者が実際に想起した内容であるため、視覚特徴の再構成だけでは不十分であり、視覚情報・意味情報・言語表現を結び付ける必要がある。

そこで本研究では、画像と言語を統合的に扱

うマルチモーダルモデルの内部表現に着目し、画像観察時の fMRI 信号と、被験者が想起した文に対応する特徴表現とを結び付けるデコーディング枠組みの構築を目的とした。特に、固定長の fMRI 信号と可変長の想起文トークン列との不整合を緩和するために、滑動平均による信号平滑化とリッジ回帰による特徴対応学習を組み合わせた。また、脳活動から直接一つの文を出力するのではなく、複数の候補文を生成し、その中から脳活動とより整合的な候補を評価・選択できるかを検証した。

[研究の内容、成果]

1. 提案手法

本研究では、画像観察時の脳活動と、被験者がその画像から心の中で想起した文との対応関係を扱うために、視覚情報と言語情報を統合的に表現できるマルチモーダルモデル LLaVA の隠れ層表現を利用した。画像観察時の fMRI 信号は固定長の時間系列として得られるのに対し、想起文を言語モデル上で表現した場合のトークン列は文ごとに長さが異なる。そのため、両者をそのまま一対一に対応づけることは難しい。この点が、本研究課題における技術的な中心問題である。

そこで本研究では、まず画像および想起文を

LLaVA に入力し、画像・言語の双方を反映した統合表現を抽出した。LLaVA は視覚情報と言語情報を同一の表現空間で扱うことができるため、画像内容と想起文内容を共通の特徴空間上で比較できる利点がある。一方、fMRI 側については、画像提示時に得られた信号に対して滑動平均を適用し、時間的なずれや局所的なノイズの影響を緩和した。これにより、固定長の脳活動系列を、可変長の言語特徴に対してより安定に対応づけることを目指した。

次に、LLaVA から得られる特徴量から脳活動を予測するエンコーディングモデルとして、リッジ回帰を用いた。リッジ回帰を採用した理由は、高次元の特徴量と多数のボクセルを同時に扱う場合でも、過学習を抑えつつ比較的安定した対応関係を学習しやすいためである。本研究では、画像または候補文から抽出した特徴量を入力とし、対応する fMRI 信号を予測するモデルを構築した。これにより、ある候補文が与えられたとき、その文に対応して予測される脳活動と、実際に観測された脳活動との近さを評価できるようにした。

さらに、脳活動から直接一つの文を生成するのではなく、画像と初期プロンプトをもとに LLaVA から複数の候補文を生成し、それらを脳活動との整合性に基づいて評価・選択する方式を採用した。

2. 実験

実験には、自然画像観察時の脳活動データを用いた。図 1 に示すように、被験者 1 名に対し、

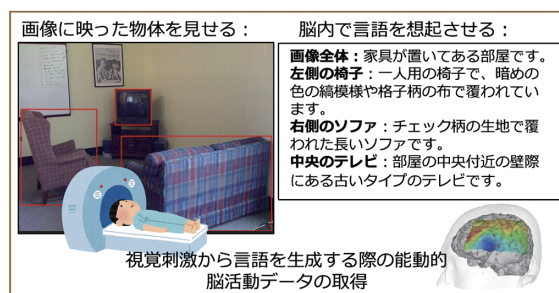


図 1 fMRI 実験全体像

COCO データセットから選出した自然画像を提示し、その際の大脳皮質 fMRI 信号を記録した。各時点におけるボクセル数は 93,452 であり、fMRI の撮像は 1TR (repetition time) = 1 秒で行われた。各試行では、まず画像全体を一定時間観察し、その後、同じ画像に含まれる主要物体へ順次注意を向けながら想起を行う構成を採用した。これにより、被験者は画像全体の場面理解だけでなく、画像を構成する個々の対象にも意識を向けながら、内容を心の中で言語化していたと考えられる。実験を通じて、合計 560 枚の画像に対応する脳活動データを収集した。このうち 448 枚を訓練用、112 枚を検証用として用いた。また、一部のブロックでは同一画像セットを再提示し、想起内容と脳活動反応の再現性も確認できるようにした。さらに、fMRI 計測終了後には各画像を再提示し、被験者自身にその画像を見てスキャナー内で想起していた内容を文章として記述してもらった。この文章を、脳活動と言語表現との対応付けに用いる教師信号として利用した。この点で、本研究が対象とするのは一般的な画像キャプション生成とは異なる。すなわち、ある画像に対して客観的にもっとも妥当な説明文を求めるのではなく、被験者本人がその画像から実際に心の中で想起した内容を扱う。

そのため、同じ画像であっても、一般的な説明文と被験者の想起文は一致しない場合があり、この個人内の意味表象をどこまで脳活動から捉えられるかが本研究の焦点となる。候補文の生成には LLaVA を用い、画像と初期プロンプトを入力として複数の文を生成した。Beam Search, Random Sampling, Top-k Sampling, Top-p Sampling など複数の手法を用いた。具体的には、Beam Search では num_beams を 4 に設定し、Sampling 系の手法では do_sample = True としたうえで、Top-k Sampling では top_k を 2, 3, 4, Top-p Sampling では top_p を 0.5, 0.7, 0.9 に設定した。これにより、生成方式の違いによって候補文の多様性や安定性がどのよ

うに変化するかを比較しつつ、脳活動との整合性という観点からどのような候補が選ばれるかを検討した。また、生成された候補文について、日本語として破綻なく出力された割合を成功率として集計し、その結果を表1に整理した。

3. 結果・考察

得られた各候補文については、LLaVA から特徴量を抽出し、あらかじめ学習したエンコーディングモデルに入力することで、その候補文に対応して予測される脳活動を算出した。そして、この予測脳活動と実測された fMRI 信号との近さを MSE や相関係数などによって評価し、どの候補文が脳活動とより整合的かを比較した。図2に示されるように、平均相関係数そのものは高いとは言えず、脳活動とマルチモーダル特徴量との対応付けはまだ十分に確立された段階にはないと考えられる。一方で、MSE の分布は比較的狭い範囲に収まっており、予測誤差のばらつきが極端に大きくないことから、少なくとも安定した基礎的対応関係は学習できていると解釈できる。したがって、本研究の手法は、画像観察時の内言的な想起内容に対して、完全な復元には至らないものの、候補文を評価するための基盤モデルとしては一定の有効性を持つと考えられる。

次に、デコーダー側の評価では、448 個の

表1 各生成方式における日本語生成の成功率

手法	成功 サンプル数	試行総数	成功率 (%)
Beam Search	1675	1792	93.5
Random Sampling	1322	1344	98.4
Top-k Sampling	1326	1344	98.7
Top-p Sampling	1314	1344	97.8

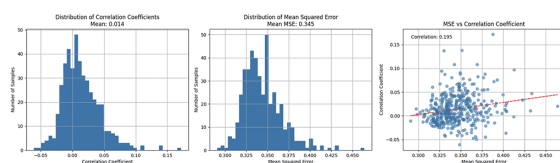


図2 エンコーディングモデルの精度：MSE と相関係数の分布および比較

fMRI サンプルに対して異なる生成方式を適用し、それぞれから得られた候補文を脳活動との整合性に基づいて評価した。生成に成功したサンプルのみを対象に MSE や相関係数を比較すると、各生成方式の値には大きな差が見られなかった（図3）。

これは、生成アルゴリズムの違い自体が脳活動との一致度を大きく左右しているというよりも、現段階では、脳活動と特徴表現との対応精度そのものが性能を制約している可能性を示している。すなわち、候補文の多様性をどれだけ増やしても、脳活動側との対応付けが十分に精密でなければ、明確な優劣は現れにくいと考えられる。

しかし、個別事例を詳しく見ると、本手法の有効性を示す重要な結果も確認された。表2では、「白い2-3人掛けで肘掛けの付いたソファ、上に黄色いクッションが置かれている」という想起文に対し、複数の生成文とそれぞれのMSE、相関係数、平均対数尤度が比較されている。この例では、「キャンディケース」や「アメリカ」など想起内容から離れた語を含む文も生成されていたが、評価指標に基づいて選ばれた文には、「ソファ」「黄色」「布地の描写」など、想起文により近い表現が含まれていた。

このことは、MSE や相関係数、平均対数尤度といった指標が、単なる信号上の近さだけではなく、ある程度は生成文の意味的適合性も反映している可能性を示している。もちろん、すべての事例で明確に想起文へ近づくわけではなく、現段階ではまだ初歩的な結果である。しか

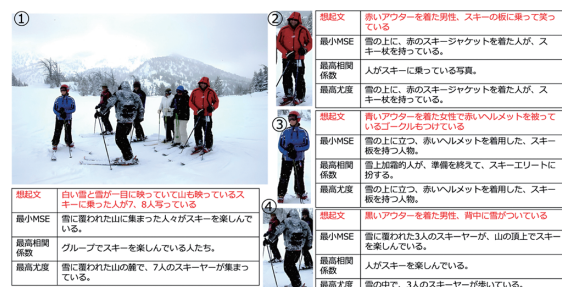


図3 生成例

表2 「想起文：白い2-3人掛けで肘掛けの付いたソファ、上に黄色いクッションが置かれている」に対する生成例

生成方法	生成文	MSE ↓	相関係数 ↑	平均対数尤度
beam_search_1	白いソファに黄色のカーペットが敷かれ、その上に色彩豊かなカーペットが敷かれています。	0.4001	0.0034	-26.6980
beam_search_2	白いソファに黄色のカーペットが敷かれ、その上に色彩豊かなカーペットが乗っています。	0.4022	0.0024	-26.6604
beam_search_3	白いソファに黄色のカーペットが敷かれ、その上に色彩豊かなカーペットが敷かれています。	0.4021	0.0110	-26.5495
beam_search_4	白いソファに黄色のカーペットが敷かれています。	0.4117	0.0022	-26.6693
random_sampling_0.2	白いソファには、柔らかなカーペットが敷かれています。	0.4140	0.0019	-26.9633
random_sampling_0.5	白いソファに黄色のシーツが敷かれ、カラフルなキャンディケースが敷かれた独特なインテリアが描かれています。	0.4033	-0.0154	-26.7185
random_sampling_0.9	アパートのリビングルームで、クリーム色のソファが見える。	0.4063	-0.0060	-26.5744
top_k_2	そのソファは、白と黄色の毛布で包まれ、複数の色の毛布で裾が飾られています。	0.4051	0.0400	-26.7571
top_k_3	白いソファの前には、床には独特のパターンがある、豊かな柔らかな雰囲気が漂っている。	0.4017	-0.0018	-26.8489
top_k_4	白いソファで囲まれた寝室。	0.4139	0.0115	-27.5148
top_p_0.5	白いソファには、黄色のカーペットが敷かれています。	0.4096	0.0150	-26.6981
top_p_0.7	リビングルームの寝室。	0.4093	-0.0425	-27.5075
top_p_0.9	和田のアメリカ屋の部屋。	0.4301	-0.0134	-27.9701

し、画像観察時の内言的な文想起という扱いの難しい課題に対して、脳活動を手がかりに候補文の妥当性を評価できることを示した点は重要である。以上より、本研究は、画像観察時の想起内容に対して、脳活動に基づいて候補文を評価・選択する基盤プロセスの可能性を示したといえる。

4. 今後の課題

本研究では、全文を生成した後に、その候補文全体と脳活動との整合性を評価する方式を採用した。しかし今後は、生成途中の各段階で脳活動情報を利用しながら文生成を制御する手法へと発展させる必要がある。たとえば、各デコード時点で脳活動との一致度を参照し、その情報を次トークンの選択に反映できれば、生成後に候補を選別するだけでなく、生成過程そのものを脳活動に沿って誘導することが可能になると考えられる。時間整合の方法についても改善の余地がある。本研究では滑動平均により固定長の fMRI 信号と可変長の言語表現との不整合

を緩和したが、より精密な時間対応モデルや、トークン長の違いを明示的に扱う方法を導入することで、対応精度の向上が期待される。また、今回のエンコーディングモデルは線形なリッジ回帰を基礎としたが、非線形モデルやより表現力の高い対応学習法を導入することで、脳活動と言語特徴の複雑な関係をより適切に捉えられる可能性がある。以上のような改善を通じて、今後はより安定した脳活動デコーディング性能を実現し、最終的には、画像観察時に人が心の中で想起する内容をより自然な文として再構成する方向へ発展させたい。さらに、その先には、発話が困難な状況にある人の意思表示支援や、人間の意味理解過程の解明といった応用的・基礎的な展開も期待される。

[成果の発表、論文など]

Liu Muxuan・西田知史・小林一郎：「心の中の言葉」はどのように予測できるか？—複数のモダリティの特徴に基づく脳活動デコーディングプロセスの構築—, 第31回言語処理学会年次大会 (NLP2025), 2025 年