深層生成モデルを用いたヒト脳活動からの動的体験再構成

高木 優
大阪大学 大学院生命機能研究科

1. はじめに

ヒトの活動の多くは視覚的な体験を伴い、その情報は幅広い脳領域で処理されている。ヒト脳内の視覚情報処理を非侵襲的に探るために、これまでの認知神経科学研究では主に機能的核磁気共鳴画像法(fMRI)を用いてきた。具体的には、被験者が fMRI 内で画像を観ている際に、機器から得られた信号値を解析することで、脳のどこにどのような視覚情報が表現されているかを明らかにしてきた。一方で、fMRI を用いた先行研究では静止画を視聴中の脳活動解析が主に行われており、より自然な動画像や音声を視聴中の脳活動を用いた解析は行われてこなかった。

以上を踏まえ、本研究では、次の 3 つの課題に取り組んだ：

1. 深層拡散生成モデルを用いた脳活動からの動画像再構成
2. 深層生成モデルを用いた脳活動からの音楽再構成（Denk and Takagi et al., arXiv 2023）。
3. 大規模言語モデルを用いた自然な動画視聴時の脳活動と意味内容との関係性調査（Nakagi,…, and Takagi, bioRxiv 2024）。

まず、1.については、fMRI 脳活動からの動的画像再構成は時間解像度やノイズの観点から難しく、静的な画像に比して大幅に精度が低い再構成となることがわかった。特にボトルネックとなったのは、動画再構成に用いるための深層拡散生成モデルでオープンソースかつ研究に用いることが可能なプレトレーニングをされたモデルが存在しなかったことである。こうしたモデルのプレトレーニングには通常数億円規模の計算資源が必要となることから、今回の研究計画の範囲で行うことは不可能であると判断した。そのため、本計画では上記の条件が整っていた 2.音楽および 3.意味的な内容を、自然な動的経験中のヒト脳活動と生成 AI との関係性を検討した結果について報告する。

2. 自然音楽視聴中の脳からの音楽再構（Denk and Takagi et al., arXiv 2023）

2.1 背景

音楽は文化を超えた普遍的なコミュニケーション手段としての役割を担っており、脳内での音楽表現は神経科学における主要な研究分野である。これまでの研究は、機能的磁気共鳴画像法(fMRI)を用いて音楽を聴く際の脳活動を捉え、脳内の音楽的特徴表現を明らかにしてきた。近年登場した音楽生成モデルは、高精度な音楽の条件付き生成が可能になっており、これによって言語的理解と実際の音楽作成の間のギャップを埋める新たな可能性が生まれている。一方で、これらのモデルの内部表現が、人間が音楽を処理する際の脳内表現とどのように関連しているのかはわかっていない。本研究では、近年我々が提案した音楽生成モデルである MusicLM と人間の脳活動との関係性を探った。

2.2 手法

本研究では、我々が過去に発表した music genre neuroimaging dataset を用いた。このデータセットでは、GTZAN データセットの 10 ジャンルに属する曲から、各ジャンルにランダムに 15 秒の音楽クリップを抽出し、刺激として用いている。これらのデータにはキャプションが付与されており、ジャンル、楽器、リズム、ムードの観点から音楽作品を記述している。

Mulan は BERT と ResNet-50 をベースとしたテキスト/音楽埋め込みモデルであり、音楽とテキストについて 128 次元の埋め込み間のコントラスト損失を最小限に抑えることを目的に学習された。MusicLM は、Transformer を利用することによって、Mulan 埋め込みをベースとして音楽を生成する条件付き音楽生成モデルである。MusicLM では、Mulan 埋め込みをより低次の音響特徴量である w2vbert トークンへと変換した後、Mulan 埋め込みへと変換し、デコーダーを使用してオーディオへと変換される。

デコーディングとは、記録された脳活動に基づいて、被験者がさらされた元の刺激を再構成する試みを指す。具体的には fMRI データに基づいて音楽埋め込みを予測し、その埋め込みに基づいて音楽を生成する。まず、5 人の参加者のそれぞれに対して fMRI 応答から刺の音楽埋め込みを予測する線型モデルを構築する。線形モデルの重みはトレーニングデータセット上で L2 正則化線形回帰を使用して推定した。予測された Mulan 埋め込みから元刺激を予測するために、既存の音楽コーパスから類似の音楽を抽出する方法と、MusicLM を用いて音楽を生成する方法の 2 つのアプローチを用いた。既存の音楽コーパスからの音楽抽出のために、Free Music Archive(FMA)データセットを用いた。音楽抽出とは異なる再構成方法として、予測された Mulan 埋め込みを用いて MusicLM モデルから音楽を直接生成した。

MusicLM の内部表現を解釈するために、MusicLM の内部表現と、fMRI により記録された脳活動との対応関係を調べた。具体的には、MusicLM の異なる音楽埋め込み(Mulan と w2vbert)を用いて fMRI 信号を予測するために、全脳ボクセルワイズエンコーディングモデルを構築した。これにより、これら 2 種類の埋め込みが脳のどの部位と対応するかの違いを探ることができる。エンコーディングモデリングでは、デコーディングとは逆に、異なる埋め込みから fMRI 応答を予測する。予測モデルの重みは L2 正則化線形回帰を用いて訓練データから推定し、その後テストデータに適用した。

2.3 結果

図 1 は、fMRI デコーディングの定量的な結果を示している。本研究では、再構成された音楽と元の刺激の類似性を評価するために、w2vbert と MuLan という異なる粒度の音楽埋め込み表現を用いて精度を定量化した。我々のアプローチは、いずれの指標でもチャンスレベルを有意に上回る制度で音楽の再構成を行うことができた。また、w2vbert に比べて MuLan の精度が高いことは、MuLan によって捉えられる音楽に関する高レベルの意味的特徴に関して、再構成された音楽がより元の刺激と似ていることを示す。

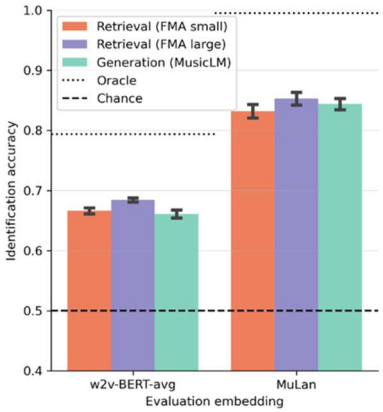

図 1：音楽再構成の定量的評価(N=5 被験者)。

図 2 は、MusicLM 内の異なるレベルの音楽埋め込み表現を用いた場合のエンコーディングモデルの予測精度を示している。いずれの埋め込み表現も、聴覚皮質を中心に高い精度で脳活動を予測した。また、MuLan 埋め込みは w2vbert よりも側頭前皮質において高い予測パフォーマンスを持つ傾向があり、MuLan が人間の脳で処理される高次元の音楽情報を捉えていることを示唆している。さらに、低次(w2vbert)から高次(MuLan)の異なるレベルの音楽埋め込み表現が、聴覚皮質内でかなり似た脳領域を予測していることがわかった。

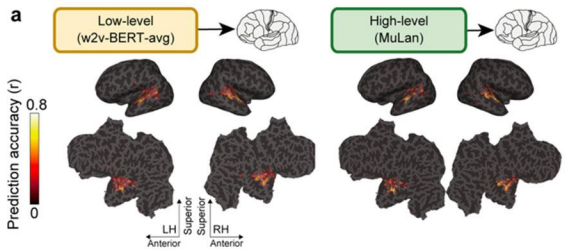

図 2：w2v-BERT と MuLan による脳活動予測。

3. 自然動画視聴中の大規模言語モデルと脳活動との対応関係の検証（Nakagi,…, and Takagi, bioRxiv 2024）

3.1 背景

近年、大規模コーパスから言語の統計的構造を学習することで自然言語の意味を理解する大規模言語モデル(LLM)が発展している。意味理解は私たち人間の知能の基盤でもあることから、意味理解に関するヒト脳活動と言語モデルの潜在表現の対応関係が注目されてきた。最近の脳研究では、言語モデルの潜在表現から特徴量を抽出し、その特徴量から脳活動を予測する脳活動エンコーディングモデルを用いて、この対応関係が探られてきた。

しかし、既存研究は意味理解の一側面に焦点を当てていた。私たちが自然な動的体験中に扱う自然言語の意味理解には、会話内容、視覚的内容、場所や時間、会話の文脈などの複数のレベルが存在する。そのため、意味理解について本来はより多くの側面を持つヒト脳と言語モデルとの対応関係を包括的に理解する上では不明な点が多く残されている。そこで本研究では、自然な動的体験中に異なるレベルでの意味理解が脳内でどのように表象されているかを、LLM と脳活動を用いて包括的に調べた。

3.2 手法

健常な日本人の被験者 6 名が 9 本の映画やドラマの DVD（10 エピソード、合計 8.3 時間）を 3 テスラの fMRI 内で自由視視聴しているときの脳活動データを集めた。本実験は、事前に実験内容を説明した上で全被験者から実験参加の書面同意を得た。実験は NICT の倫理・安全委員会から承認を得た。fMRI データの前処理には、運動補正と解像度補正とトレンド除去を行った。

刺激動画として、複数の詳細な言語アノテーション付きの 9 本の映画やドラマの DVD（10 エピソード、合計 8.3 時間）を刺激動画として使用した。映画とドラマは様々なジャンルを含んでおり、8 本が海外の映画またはドラマ、1 本が日本のアニメーションである。言語アノテーションデータは、音声内容の書き起こし(Speech)、シーンごとの物体情報(Object)、シーンごとの物語内容(Story)、物語のあらすじ(Summary)、および時間と場所の情報(TimePlace)の 5 種類から成る。それぞれの意味内容は自然言語によって刺激動画を説明するアノテーションであるが、その説明内容の性質やアノテーションのタイムスパンが異なっている。具体的には、Speech は発話ごとに、Object は 1 秒ごとに、Story は 5 秒ごとに、Summary は 1〜2 分ごとに、TimePlace は画面展開ごとにアノテーションされている。各意味内容が持つ潜在的な表現を刺激特徴量として抽出するために、我々は各言語アノテーションを入力として 5 種類の言語モデルの単語埋め込みまたは潜在表現を取得した。

異なるレベルの意味内容がヒト脳においてどのように表現されているのかを調べるために、5 種類の意味内容に関する言語モデルの潜在表現から脳活動を予測する脳活動エンコーディングモデルを構築した。モデルの構築方法は音楽生成 AI の研究と同様である。モデルの重み推定には、訓練データに対して L2 正則化付き線形回帰を使用し、その後テストデータに適用した。

3.3 結果

図 3 に、異なるレベルの意味内容から抽出したそれぞれの特徴量が脳活動をどのように説明するのかを調べた結果を示す。図 3 より、Speech、Object、Story の意味内容に関する特徴量からの予

測精度は、Summary、TimePlace の意味内容に関する特徴量からの予測精度より高いことがわかった。また、Speech、Object、Story について、従来型言語モデルの Word2Vec と比較して、より大規模な LLMs（特に Llama2）の方が高い予測精度を持つことが被験者間共通で確認された。この結果は、特に Story において顕著であった。

図 4 は、5 種類の意味内容について Llama2 の潜在表現から脳活動を同時に並列して予測した時の大脳皮質における予測精度である。図 3 より、構築したエンコーディングモデルは大脳皮質上の幅広い領域（視覚・聴覚に関わる脳領域～より高次の脳領域）の脳活動を予測できていることが確認できた。

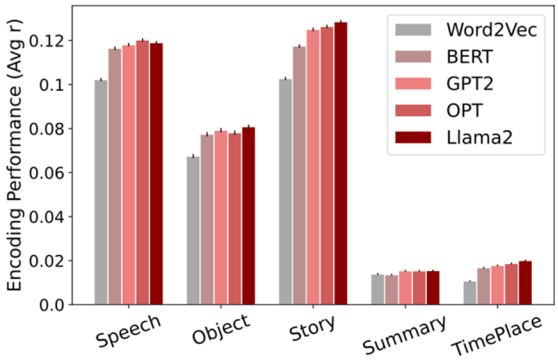

図 3：5 種類の意味内容に関する言語モデルの潜在表現からの脳活動予測精度。

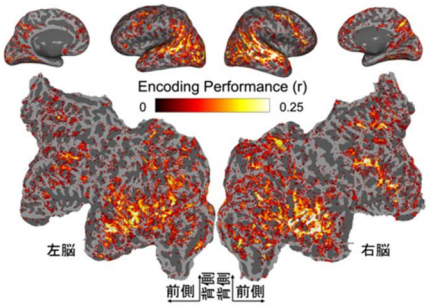

図 4：5 種類の意味内容に関する Llama2 の潜在表現から脳活動を同時に予測したときの予測精度を大脳皮質上にマッピングした結果。

4. おわりに

本研究では、自然な動的体験中のヒト脳活動と生成 AI の潜在表現との対応関係を調査した。

まず、音楽生成 AI を用いた研究では、人間の脳活動から音楽を再構成するために、fMRI データから予測された音楽埋め込みに MusicLM を条件付けることにより、意味レベルで元の音楽刺激に似た音楽を生成することができた。また、エンコーディングモデルを構築することにより、テキスト音楽先生モデルと人間の脳との関係を定量的に評価した。具体的には、音楽の高次元の意味情報と低次元の音響特徴が人間の脳でどこに、どの程度表現されているかを評価した。テキストから音楽へのモデルは急速に発展しているが、その内部プロセスはまだ十分に理解されていない。本研究で使用された音楽生成モデルが、これまでのディープラーニングモデルと比べて脳にインスパイアされたものではないにもかかわらず、このような脳との対応関係が生じたことは興味深い。

次に、LLM を用いた研究では、異なる意味内容に関する LLMs の潜在表現とヒト脳活動の対応関係を定量的に比較した。結果、従来型の言語モデルと比較して、LLMs の潜在表現は、特に高次の言語表現を持つ物語内容に関する脳活動をよく説明することがわかった。加えて、異なる意味内容に関する LLMs の潜在表現は、それぞれ異なる脳領域の脳活動を固有に説明することもわかった。

参考文献

(1) Denk et al., arXiv, (2024).

(2) Nakagi, et al., bioRxiv., (2024).