jh250088

脳活動予測に特化した大規模言語モデルの構築

小林一郎（お茶の水女子大学）

概要

本申請研究では，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，言語モデルから得られた特徴量を入力情報として用い，構築された符号化モデルで脳内活動状態の予測を従来のモデルに対して精度の向上を実現することを目的とする．とくに，脳活動データと言語の特徴量の対応関係を学習する Transformer ベースのモデルに入力する際の脳活動データの抽出方法が脳内状態予測精度に与える影響について検証を行った．脳状態の抽出方法に 3 つの手法を用いて実験を行った結果，ピアソンの積率相関係数のボクセル平均は，脳活動データと言語特徴量の対応関係を事前に学習したモデルの精度よりも事前学習済み BERT の特徴量を用いるベースラインが 0.197 と最も高くなり，先行研究の知見と一致しなかった．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，実験設定に課題が残されていることが示唆された．

1　共同研究に関する情報

1.1　共同研究を実施した拠点名

- 東京大学 情報基盤センター

1.2　課題分野

- データ科学・データ利活用課題分野

1.3　参加研究者の役割分担

- 小林一郎（研究総括）
- 宮尾祐介（共同研究者）大規模言語モデルの学習についてアドバイスをする．
- LUO Ying（共同研究者）モデル学習の枠組みを提案．
- LIU Muxuan（共同研究者）脳活動データの処理および符号化モデルの調査．
- 田屋侑希（共同研究者）事前学習済みモデルの活用法の検討．
- 濵田 愛（共同研究者）モデルの学習および検証実験．

2　研究の目的と意義

深層学習はそれまでの知能情報処理に大きな影響を与え，それは様々な分野に導入され新しい知見が日々生産されている．脳神経科学の分野においても，2014 年に Yamins ら [1] によってヒト大脳皮質における階層情報表現と深層学習の階層情報表現に相同性があることが示され，その知見は基礎・応用両面に渡り多くの発展可能性を示唆し，大脳皮質の動作原理の一つとして深層学習が有効であること [2]，深層学習の援用により脳活動からの認知内容解読が高度化できること [3] 等が実証されて以来，深層学習のモデルを作業モデルとして，ヒト脳内のメカニズムを解明する，あるいは，ヒト脳内の情報を読み解く解読手法の開発が進められてい

る．脳に与えられる刺激から脳内状態を予測するためには，刺激となる情報からその特徴量を抽出し，その特徴量から脳活動を予測する（あるいは脳活動から特徴量を予測）回帰モデル（「符号化モデル [4]」と呼ばれる）を構築する必要がある．ヒト脳に与えられる刺激は，主に，画像, 言語, 音, 等であり，それらの特徴量を抽出するために，深層学習モデルを用い，モデルの中間層に表現される状態をその刺激の特徴量として利用する．これまでに畳込みニューラルネットワークモデルを用いて，視覚刺激に対する特徴量の抽出を行い，その特徴量から構築した符号化モデルを用いて，ヒト脳活動の状態を予測する研究が進められ，多くの知見を得ている [5][6]．一方, 言語刺激に対しても事前学習済み言語モデル BERT[7] や GPT-2[8] などが出現して以来，それらのモデルの中間層から得られる特徴量を言語刺激に対する特徴量として，ヒト脳内の状態を予測する研究が進められている [9]．また，2023 年には Huth らのグループにおいて，GPT[10] を用いて逐次的に予測する単語をヒト脳活動データから符号化モデルを通じて回帰を行い予測した単語と一致率が高いものを選択していくことで文章を生成する，ヒト脳活動を自然言語で解読する手法を開発している [11]．これらの研究において，言語刺激からヒト脳内状態を予測する符号化モデルを構築する際に，事前学習済み言語モデルを言語刺激の特徴量として使用されている．この事前学習済み言語モデルを他のモダリティと融合させ，言語とそのモダリティ間の変換精度を高める手法として，音声とテキスト（言語）間の対応関係を事前に学習させた SpeechBERT[12] や画像と言語間の対応関係を事前に学習させた VL-BERT[13] などが提案されている．それと同様に，申請者のグループにおいては，脳活動データと言語の対応関係を事前に学習させた汎用言語モデル BrainBERT [14] を開発し，言語刺激を BrainBERT の埋込みベクトルを用いて，ヒト脳活動を予測した際の精度において SOTA を実現した．その後も BrainBERT に用いる事前学習済み言語モデルを様々なものに変更し予測精度の検証を進め，BrainLM（Brain Language Model）という名称に変更し，研究成果をまとめている [15]．このようにヒト脳状態と言語刺激の特徴量に対して対応関係をとった事前学習済み言語モデルが存在することは，外部刺激からより正確にヒト脳内状態を予測でき，かつ，ヒト脳活動データから外部刺激を予測するという基本的な部分において性能を高めた符号化モデルを構築できることになり，脳神経科学，Brain-Machine Interface の開発，脳活動を模倣した人工神経回路網モデルの構築，など多くの分野に影響を与えることができる．これまで言語の刺激を抽出していた事前学習済み言語モデルは，現在，大きな発展をとげ，様々な大規模言語モデルの出現に至っている．また，申請者のグループの研究において，大規模言語モデルを符号化モデルに用いることでヒト脳活動の予測精度が上がることをすでに確認している [16]．このような背景から，本申請研究では，計算機能力の限界によって，これまで比較的小さな汎用言語モデルでしか BrainLM を構築できなかったものを，大規模言語モデルから抽出する言語特徴量を使ったものに変更し，BrainLM をさらに言語，視覚情報などと脳活動を強く結びつけた事前学習済み言語モデルとして構築することで，それを資源として使用可能とし，脳神経科学，MBI の開発，医療，情報科学など様々な分野の研究への貢献を目指す．

3 当拠点公募型研究として実施した意義

大規模言語モデルの学習には大きな計算リソースが必要となる．本公募型研究において，そのための計算機環境を利用することができた．

4 前年度までに得られた研究成果の概要

該当なし．

5 今年度の研究成果の詳細

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（5.1 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（5.2 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．

5.1 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．

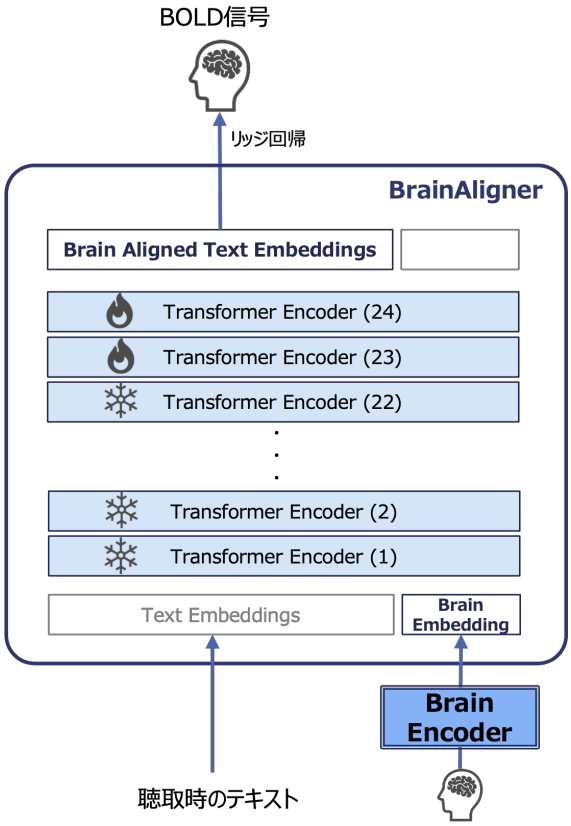

図 1 言語刺激と脳活動の対応関係を学習する言語モデルおよびそれを用いた脳内状態予測の概要図．学習時のみ Brain Encoder を活用して脳状態を入力し，評価時は聴取時のテキストのみ入力．

5.1.1 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 [17] 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．

5.1.2 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) [18] 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 を用いて正規化し，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．

5.1.3 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．

5.2 BrainAligner

BrainAligner は, 事前学習済み BERT-large-uncased を初期重みとし，5.1 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 [19](5.4.1 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 [7] 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．

5.3 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 $\{0.01, 0.1, 1, 10, 100, 1000\}$).

5.4 実験

5.4.1 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時の fMRI 脳活動データ [19] 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．

5.4.2 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（5.1.1 節）, High_EV（5.1.2 節）, ES（5.1.3 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．

5.4.3 実験結果

表 1 に，3 種類の 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，符号化モデルで各ボクセルの脳内状態予測をし，実測値とのピアソンの積率相関係数 r を計算した結果を示す．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，次いで BrainLM が近い値を示した．一方，ES および High_EV は相関が低く，Baseline・BrainLM との差が確認された．

さらに，Brain Encoder として脳活動事前学習モデル Brain LM を用いた Brain Aligner について，MLM 学習におけるマスク率を系統的に変えて学習させた．得られた特徴量をもとに符号化モデルで脳内状態予測を行い，実測値との相関係数を計算した結果を表 2 に示す．

5.5 考察とまとめ

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

表 1　脳内状態予測結果を元に算出した相関係数の平均値と中央値.

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

表 2　Brain Encoder として Brain LM を用いた Brain Aligner の脳内状態予測結果を元に算出した相関係数の平均値と中央値.

手法	被験者	r_mean	r_median
4*maskp030	UTS01	0.194345	0.193985
	UTS02	0.193652	0.193170
	UTS03	0.195495	0.194219
	平均	**0.194497**	**0.193791**
4*maskp045	UTS01	0.194185	0.193814
	UTS02	0.193517	0.193003
	UTS03	0.195228	0.194034
	平均	0.194310	0.193617
4*maskp060	UTS01	0.193843	0.193490
	UTS02	0.193572	0.193136
	UTS03	0.195386	0.194133
	平均	0.194267	0.193586
4*maskp075	UTS01	0.194205	0.193793
	UTS02	0.193625	0.193176
	UTS03	0.195241	0.194030
	平均	0.194357	0.193666
4*maskp090	UTS01	0.193931	0.193580
	UTS02	0.193468	0.193081
	UTS03	0.194802	0.193675
	平均	0.194067	0.193445

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．また，BrainAligner において MLM のマスク率を変化させた場合，いずれの設定でも平均相関係数は近い値を示しつつ，マスク率 30% が最も高い結果となった（表 2）．差分は小さいことから，少なくとも本実験範囲では，マスク率による影響は限定的であり，極端に高いマスク率では文脈復元が難しくなる一方，低すぎるマスク率では学習信号が弱くなる可能性がある，という一般的なトレードオフと整合的な範囲にとどまったと解釈できる．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　進捗状況の自己評価と今後の展望

初めて他大学の情報基盤センターの計算機資源を使わせてもらったため，当初，GPU 利用方法に手間取ってしまったため，あまりたくさ

んの実験ができず申請した計算機使用費を大幅に下回った利用となった．今後は，もっと計画的に実験準備を進め，より大きな言語モデルの学習を行えるようにするつもりである．

参考文献

[1] Daniel L. K. Yamins, Ha Hong, Charles F. Cadieu, Ethan A. Solomon, and James J. Seibert, Darren an d DiCarlo. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, Vol. 111, No. 23, pp. 8619–8624, 2014.

[2] Nikolaus Kriegeskorte. Deep neural networks: A new framework for modeling biological vision and brain information processing. *Annual Review of Vision Science*, Vol. 1, pp. 417–446, 2015.

[3] Eri Matsuo, Ichiro Kobayashi, Shinji Nishimoto, Satoshi Nishida, and Hideki Asoh. Generating natural language descriptions for semantic representations of human brain activity. In He He, Tao Lei, and Will Roberts, editors, *Proceedings of the ACL 2016 Student Research Workshop*, pp. 22–29, Berlin, Germany, August 2016. Association for Computational Linguistics.

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

[5] Radoslaw Martin Cichy, Aditya Khosla, Dimitrios Pantazis, and Aude Torralba, Antonio an d Oliva. Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual obj ect recognition reveals hierarchical correspondence. *Scientific Reports*, Vol. 6, No. 1, p. 27755, 2016.

[6] Michael Eickenberg, Alexandre Gramfort, Gaël Varoquaux, and Bertrand Thirion. Seeing it all: Convolutional network layers map the function of the human visual system. *Neuroimage*, Vol. 152, pp. 184–194, May 2017.

[7] 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.

[8] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI blog*, Vol. 1, No. 8, 2019.

[9] Jingyuan Sun, Shaonan Wang, Jiajun Zhang, and Chengqing Zong. Neural encoding and decoding with distributed sentence representations. *IEEE Trans. Neural Networks Learn. Syst.*, Vol. 32, No. 2, pp. 589–603, 2021.

[10] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. 2018.

[11] 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.

[12] Yung-Sung Chuang, Chi-Liang Liu, Hung yi Lee, and Lin shan Lee. SpeechBERT: An Audio-and-Text Jointly Learned Language Model for End-to-End Spoken Question Answering. In *Proc. Interspeech 2020*, pp. 436–440, 2020.

[13] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. Vl-bert: Pre-training of generic visual-linguistic representations. In *International Conference on Learning Representations (ICLR)*, 2020.

[14] Ying Luo and Ichiro Kobayashi. Brainbert: A language model capturing the correspondence between brain activities and language. In *The 22nd International Symposium on Advanced Intelligent Systems (ISIS2021)*, 2021.

[15] Yin Luo and Ichiro Kobayashi. Brainlm:

Enhancing brain encoding and decoding capabilities with applications in multilingual learning. *Journal of Advanced Computational Intelligence and Intelligent Informatics*, Vol. 29, No. 4, pp. 754–761, 2025.

[16] Anna Sato and Ichiro Kobayashi. Decoding semantic representations in the brain under language stimuli with large language models. In *COLING Workshops*, 2025.

[17] 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.

[18] 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.

[19] 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.