2K3-OS-33a-2in1

深層学習による画像刺激時のfMRI脳活動データからの文生成

Natural Language Generation with fMRI Data of Brain Activity Evoked by Visual Stimuli using Deep Learning

松尾 映里 *1 Eri Matsuo　小林 一郎 *1 Ichiro Kobayashi　西本 伸志 *2 Shinji Nishimoto　西田 知史 *2 Satoshi Nishida　麻生 英樹 *3 Hideki Asoh

*1お茶の水女子大学 Ochanomizu University　*2情報通信研究機構 脳情報通信融合研究センター National Institute of Information and Communications Technology

*3産業技術総合研究所 人工知能研究センター National Institute of Advanced Industrial Science and Technology

Quantitative analysis of human brain activity based on language representations, such as semantic categories of words, has been actively studied in brain and neuroscience. This study attempts to generate natural language descriptions for human brain activation phenomena evoked by visual stimuli by employing deep learning. Due to the lack of brain training data, the proposed method employs a pre-trained image-captioning system using the encoder-decoder architecture. To apply brain activity data to the image-captioning model, we train a model to learn the corresponding relationships between brain activity data and image features. The results demonstrate that the proposed model can recognize semantic information of human brain activity and generate description using natural language sentences. We conducted experiments with data from brain regions known to process visual stimuli. The results suggest that semantic processing of visual stimuli is performed using the entire cerebral cortex.

1. はじめに

近年，脳神経生理学において，人間の脳内で処理される意味情報表現を定量的に分析する研究が盛んになっている．なかでも，言語処理分野における機械学習手法の発展に伴い，単語意味カテゴリなどの言語表象に基づいて脳における意味表象の解釈やモデル化を行う研究が増えている．本研究は，人が画像刺激によって頭の中に抱いた意味表象，すなわち画像が人に想起させる事象を，fMRI で観測した脳活動データを用いて，自然言語文によって説明する深層学習手法を構築する．しかし，fMRI による観測はコストが大きく，また個々人の持つ脳の形が異なるため，大量の学習データを要する深層学習を十分に行うための大規模なデータ収集は困難である．そのため，事前に訓練された画像に映る事象を言葉で説明するキャプション付け手法を援用することで少量データの効果的利活用を行った．

実験では 3 通りの設定で提案モデルを学習させ，その結果を比較した．次に，全脳情報の代わりに画像処理を行う脳領域のみを入力としたときの精度の変化を調査した．

2. 関連研究

脳神経活動データから人が想起している意味情報を解析する手法は，複数の先行研究において，脳活動データと言語意味表象の対応関係を捉えることで実現されており，なかでも統計的言語モデルが脳活動における感覚や文脈の情報に伴う高次の表象表現を説明するのに適したモデルであることが指摘されてきた [Nishimoto 11, Huth 12, Cukur 13]．Cukur ら [Cukur 13] は，動画像中の物体に注目し認識する際に，どのように認識の意味形態が変化するかを脳活動データから推定している．Huth ら [Huth 12, Huth 16] は，動画像中に現れる物体や動作を類義語体系である WordNet の語彙で表現して動画像の刺激と脳神経活動との関係について調査し，脳の皮質における意味のマップを作成するとともに，皮質の広い領域にまたがる多様な表象によって意味情報が表現されることを示した．しかし，これらの先行研究では単なるラベル分類に基づき単語のみを意味表象の推定を対象とした分析しか行われていない．本研究では，文生成モデルとして有効な手法である深層学習を用いて，より説明力の高い自然言語説明文を出力することで，脳活動の更なる定量的理解を目指す．

連絡先: 松尾映里，お茶の水女子大学大学院 人間文化創成科学研究科 理学専攻 情報科学コース 小林研究室，〒 112-8610 東京都文京区大塚 2-1-1，03-5978-5708，matsuo.eri@is.ocha.ac.jp

3. 提案手法

本提案手法は，3.1 節および 3.2 節に示す 2 種類のモデルを組み合わせることで fMRI により観測された脳活動データを入力とし，そのとき人が想起している内容を説明する自然言語文章の生成を目指す．図 1 に概要を示す．

3.1 (A) 画像→説明文モデル

本提案モデルの主部分として，深層ネットワークの枠組みである Encoder-Decoder Network（Enc-DecNet）[Cho 15] を用いて実装される画像キャプション付けモデルを用いる [Vinyals 15]．従来，画像説明文の生成に対しては，既存の説明文を検索しランキングする手法．あるいは画像から抽出される特徴量を元にテンプレートを埋める手法が主に用いられていたが，近年，機械翻訳において sequence-to-sequence モデル [Sutskever 14] として知られる，Enc-DecNet に基づく研究が数多く報告されている．Enc-DecNet では，Encoder, Decoder の役割を果たす 2 つの深層学習モデルを組み合わせることで入力を中間表現に変換（encode）し，再び復号（decode）して別の形に出力する．画像説明文生成においては，画像特徴量を中間表現とし，CNN による画像特徴抽出と LSTM による画像特徴量を用いた文生成から構成されるものが多く，本研究でも VGGNet[Simonyan 15] を Encoder，2 層 LSTM-LM[Cho 15] を Decoder とした画像→画像特徴量→説明文モデルを採用する．

3.2 (B) 脳活動データ→画像特徴量モデル

上記のプロセスを脳活動データに適用するため，fMRI 脳活動データを入力として，被験者に与えた画像から VGGNet により抽出される画像特徴量を予測するモデル，すなわち脳活動

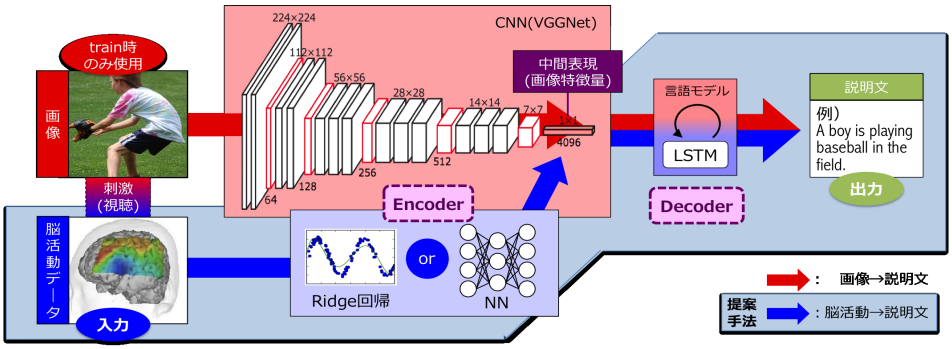


図 1: 本研究の概要

データを上記 (A) 画像→説明文モデルにおける中間表現に変換するモデルを用意した．実験では Ridge 回帰，3 層 Neural Network（NN），5 層 Deep NN（DNN）の 3 通りの実装を行い，どの手法が学習モデルとして適切かを比較調査した．また，学習データ量不足による学習の遅延や過学習を防ぐため，DNN については積層 AutoEncoder[Bengio 08] による事前学習を行った．

4. 処理の流れ

step 1. 脳活動情報の中間表現への変換
(B) 脳活動データ→画像特徴量モデルを用いて，画像視聴時の脳活動データから，そのとき見ている画像から VGGNet によって得られる画像特徴量を予測する．以降の Step では，この特徴量を中間表現として (A) 画像→説明文モデルの処理を行う．

step 2-1. LSTM-LM による単語予測
step 1 において算出された画像特徴量と 1 時刻前の LSTM の隠れ状態を入力として，LSTM-LM で単語を出力．

step 2-2. 単語出力の反復による文生成
文末記号が出力されるか設定した最大文長を超えるまで step 2-1 を繰り返し，1 語ずつ出力して文章を生成．

このように (A) 画像→画像特徴量→説明文モデル，(B) 脳活動データ→画像特徴量モデルを学習させ，順次実行することで，(C) 脳活動データ→画像特徴量→説明文モデルを実現する．

5. 実験

実装の際，深層学習のフレームワーク Chainer[*1] を利用した．

5.1 実験 (A)：画像→説明文モデル

5.1.1 実験設定

学習のためのデータセットとして，414,113 ペアの静止画とその説明文からなる Microsoft COCO[*2] を使用する．ハイパーパラメータの設定については，画像説明文生成モデルの先行研究に基づいて調整した [Vinyals 15, Mitchell 16]．学習するパラメータは標準正規分布乱数により初期化したが，単語を特徴表現空間に写像する word embedding 層は Skipgram を用いて事前学習した word2vec[Mikolov 13] を初期値とし，VGGNet は事前学習したものを用いて更新を行わない．学習に関する詳細設定は表 1 の最左列に示す．

5.1.2 実験結果と考察

epoch 毎に test 用画像からの出力の perplexity を記録し，その値により学習の進度を確認した．また，test 用画像から無作為に選んだ画像に対して生成した説明文を表 2 に示す．

1 例目は十分に妥当な説明文が生成され，2 例目も色を含め主語を正確に捉えられている．また，文章全体に大きな崩れはなく細部の前置詞（in,on）や冠詞（a,an）の区別も正しくついており，出力された説明文は内容的にも文法的にも画像の大意を認識し表現できていると言える．このように学習に使われていない画像に対しても相応な説明文を生成でき，かつ perplexity も約 2.5 で収束していることから画像→説明文モデルについては適切な学習が進んだと評価できる．また，1epoch 未満の学習のみを行った結果 [松尾 16] と比較すると評価指標も生成文も大きく精度を向上させており，深層学習のためには十分な学習量が必要であることが確認できる．

また，生成文に見られる誤りの傾向としては，2 例目にあるような立っている状態を sitting，洗面台を toilet と表すなどの言語処理より画像認識処理に依存したものと考えられる細部の単語誤りが多く，文法誤りはほとんど見受けられなかった．

5.2 実験 (B-1)：脳活動データ→画像特徴量モデル

5.2.1 実験設定

脳活動と画像特徴量の対応関係を学習するためのデータセットとして，自然動画像を一名の被験者に見せた時の血中酸素濃度依存性信号（BOLD 信号）を functional Magnetic ResonanceImaging（fMRI）を用いて記録した脳神経活動データ [Nishimoto 11]，および fMRI のデータ収集と同期して動画像から切り出したフレーム（静止画）を使用する．刺激として用いる自然動画像は映画，自然，人工物，人間，3D アニメなどの十数秒の様々な種類の動画が混ざっている．立体撮像 96×96×72 voxel のうち皮質に相当する 65,665 次元分のデータ列を入力とし，その時見ている画像から VGGNet により得られた 4,096 次元の画像特徴量との対応を学習する．train 用データ数は 4,500（2 秒毎に 9,000 秒分記録）であり，直接多層 NN を学習するには少量となる．学習するパラメータは Ridge 回帰，3 層 NN については標準正規分布乱数により初期化したが，5 層 DNN についてはデータ不足による過学習を回避し学習を早めるため，対応する画像のない脳活動データを用いて AutoEncoder による事前学習を行って得られた重みを初期値とした．学習に関する詳細設定は表 1 の右 3 列に示す．

5.2.2 実験結果と考察

epoch 毎に平均二乗誤差を記録し，3 種モデルがいずれも収束したことを確認した．数値上は Ridge 回帰（約 1.04），5 層 DNN（約 1.11），3 層 NN（約 1.13）の順に性能が高いが，ほとんど近い値に収束しているため，次節の実験 (C-1) にて 3 種モデルの比較・考察を行う．

5.3 実験 (C-1)：脳活動データ→説明文モデル

5.3.1 実験設定

実験 (A)，(B-1) で学習した画像→説明文モデルと脳活動→画像特徴量モデル 3 種を組み合わせ，脳活動データからの説明文生成を 3 通り実行し，同時にその時見ている画像から直接 (A) 画像→説明文モデルによる説明文生成も行った．

*1 http://chainer.org/
*2 http://mscoco.org/

表 1: 実験設定（詳細）

	(A) 画像→画像特徴量→説明文モデル	(B) 脳活動データ→画像特徴量モデル		
		Ridge 回帰	3 層 NN	5 層 DNN
データセット	Microsoft COCO	動画刺激による脳活動データ		
学習量	414,113 sample×100 epoch	4,500 sample×1,000 epoch		
アルゴリズム	Adam	Ridge regression	stochastic gradient descent	
学習に関するハイパーパラメータ	a=0.001, b1=0.9, b2=0.999, eps=1e-8 勾配閾値：1 L2 正則化項：0.005	L2 正則化項：0.5	学習率：0.01 勾配閾値：1 L2 正則化項：0.005	
学習するパラメータの初期値	word embedding：word2vec VGGNet：事前学習済み・学習せず それ以外：標準正規分布乱数	標準正規分布乱数	標準正規分布乱数	教師なし脳活動データを用いた AutoEncoder による事前学習（7,540sample×200epoch）
層ユニット数	各層 512	65,665 - 4,096	65,665 - 8,000 - 4,096	65,665 - 7,500 - 6,500 - 5,500 - 4,096
語彙	頻出語 3,469 語	ー		
誤差関数	交差エントロピー	平均二乗誤差		

表 2: 画像から生成した説明文の例，画像はランダムに選出

A man is surfing in the ocean on his surfboard.

A black and white cat is sitting on the toilet.

5.3.2 実験結果と考察

train 用データから 2 つ，test 用データから 2 つ選んだ 4 つの脳活動データに対して生成した説明文およびその時の画像，さらに比較のため (A) 画像→説明文モデルによる説明文を表 3 に示す．提案手法により，脳活動データのみを入力として，人間が解釈しうる説明文章が生成されることが確認された．実験 (A) のモデルと同様，前置詞や冠詞など含め安定して正しい文法表現の名詞句または文として出力された．また，Ridge 回帰，3 層 NN を用いたモデルについては，train データに対しての脳活動データからの生成文と画像からの生成文が一致していることから，実験 (B-1) による脳活動→画像特徴量への対応が学習できており，提案手法が機能していることが確認できる．しかし，5 層 DNN ではどのような入力に対しても全く同じ文が出力され．test データに対しては Ridge や 3 層 NN でも生成文の精度は低くなった．これは，学習すべきパラメータ数に比べて入力次元（65665 次元）が大きく，train データ数（4,500sample）が少ないこと，あるいはハイパーパラメータの調整不足などによる過学習が原因と考えられる．なお，AutoEncoder による事前学習を用いず，初期値をランダムに設定した 5 層 DNN を用いても，同様に同じ文が出力される過学習に陥った．また，実験 (A) において画像→説明文モデルは適切に学習されたと結論付けたにも関わらず，画像から直接生成された文も画像の内容を捉えられているとは言い難い．これは，画像の「質」が異なることが原因の一つと考えられる．Microsoft COCO の画像は画像認識処理のために用意されたデータセットであり，内容がわかりやすく，明快な説明文を付与しやすいものが多い．一方，脳活動に与えた自然動画刺激は，ぶれ，暗転，文字，見切れ，アニメなどを含んだ様々な種類の動画の 1 カットであり，人間でも自然言語文による説明が困難なものが混じっている．したがって，データを増やしての実験や，説明文の付与が容易な画像刺激を用いての追加実験が望まれる．また，興味深いのは，train の 2 例目では画像より脳活動を用いた方が適切な説明文が生成されている点である．人間の脳活動において時計とはさみを混同するとは考えにくく，VGGNet での画像認識処理における誤りであると推測されるため，脳活動を介して得られた画像特徴量の方がより有効に働いたのではないかと考察できる．

5.4 実験 (B-2)：脳活動データ→画像特徴量モデル

実験 (C-1) では，実験 (B-1) により学習されたモデルが過学習してしまっており，その原因として対応関係を学習するには脳活動データが少なすぎる，かつ高次元すぎることが考えられることを示した．そこで，脳全体を用いるのではなく，脳のうち画像刺激により反応する部位のみを入力として使用することでデータの次元を落として (B) 脳活動データ→画像特徴量モデルを学習させることで精度の向上を狙う．

5.4.1 実験設定

学習のためのデータセットとして，Exp.(B-1) で用いた自然動画像刺激によって引き起こされた脳皮質の神経活動データの 65,665 個の voxel のうち，画像処理に用いられる領域を選び入力とする．西田ら [Nishida 15] らは動画刺激による脳活動データと動画の内容を表す単語の word2vec による分散意味表現の対応関係を学習するモデルを構築し，単語意味表現から予測した脳活動データと実測した脳活動データとの相関係数を用いて，各 voxel が予測にどれだけ寄与したかを示す予測精度を求めた．本研究においては予測精度 c.c.=0.05，0.1，0.15，0.2 を閾値として，画像刺激下の脳活動のうち意味抽出に使われる voxel，すなわち画像処理に有用な voxel を選択した．それぞれの voxel 数は 21,437，9,923，5,961，3,539 となった．反対に，さらに高次元な入力として大脳皮質以外に主に記憶や空間情報に関係する皮質下部を加えた 89,206 次元のデータについても実験を行った．いずれの設定でも，実験 (C-1) において最も生成文の質が高かった 3 層 NN を学習モデルとして用い，他の実験設定は実験 (B-1) と同様とする．

5.4.2 実験結果と考察

epoch 毎に test 用データに対する平均二乗誤差を記録し，その減少により学習の進度を確認した．89206 次元および 21437 次元データを用いた実験では，評価指標は実験 (B-1) でも 3 層 NN として示した 65665 次元データを用いた結果と同様にほぼ収束しているものの，その値はそれぞれ約 1.33，1.16 となり約 1.11 より若干悪化した．一方で，9923 次元，5961 次元，3539 次元のデータを用いた実験では，評価指標から過学習が発生していることが示された．それぞれ誤差の最小値は約 1.17(24epoch)，1.13(30epoch)，1.10(30epoch) となり，3539 次元データについては数値上は最も良い学習結果となった．次節の実験 (C-2) にて 6 種モデルの具体的な比較・考察を行う．

5.5 実験 (C-2)：脳活動データ→説明文モデル

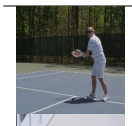
5.5.1 実験設定

実験 (A) で学習した画像→説明文モデルと実験 (B-2) で学習した脳活動→画像特徴量モデル 6 種を組み合わせ，脳活動データからの説明文生成を 6 通り実行した．過学習が確認された 9923 次元，5961 次元，3539 次元の 3 種モデルについては，学習を早期に打ち切り，最も誤差が小さかった epoch 時点のパラメータを使用した．

表 3: 実験 (C-1)：被験者が見ていた画像，その時の脳活動から生成した説明文 3 通り，画像から生成した説明文の例

		Ridge 回帰	3 層 NN	5 層 DNN	Image → Caption Model
train data		A group of people walking down the street.	A group of people walking down the street.	A fire hydrant sitting on the side of an empty street.	A group of people walking down the street.
		A pair of scissors sitting on the ground.	A close up of an orange and white clock.	A fire hydrant sitting on the side of an empty street.	A pair of scissors sitting on the ground.
test data		A man standing in front of an airplane.	A bench sitting in the middle of some grass.	A fire hydrant sitting on the side of an empty street.	A herd of sheep standing in the grass.
		A bird sitting on the branch of an apple.	A bird sitting in top of an orange tree.	A fire hydrant sitting on the side of an empty street.	A bird is perched on top of the tree branch.

表 4: 実験 (C-2)：被験者が見ていた画像と，その時の脳活動から生成した説明文 6 通り

	3,538 voxels ($c.c. > 0.2$)	5,961 voxels ($c.c. > 0.15$)	9,923 voxels ($c.c. > 0.1$)	21,437 voxels ($c.c. > 0.05$)	65,665 voxels (all cortex)	89,206 voxels (+subcortex)
	A young man is doing tricks on his skateboard.	A man is playing tennis on the court.	A young man is playing tennis on the court.	A man is playing tennis on the court.	A man is playing tennis on the court with his racket.	A man is playing tennis on the court with his racket.
	A man sitting on the ground with an umbrella.	A polar bear is standing in the water.	A dog laying on the ground next to an orange frisbee.	A dog laying on the ground next to an orange frisbee.	A black and white dog laying on the ground.	A dog is sitting on the floor in front of an open door.

5.5.2 実験結果と考察

train 用データから選んだ 2 つの脳活動データに対して生成した 6 つの説明文およびその時の画像を表 4 に示す．予想に反し，高次元なデータを用いた学習モデルの方がより適切な説明文を出力する傾向が見られた．2 例目では細部は間違えているが 2 匹の犬の色を認識できており，65,665 次元の皮質全体を用いた元のモデルが最も良く学習されたと考えられる．反対に，低次元データを用いたモデルからは内容にそぐわない不適切な文が生成された．100epoch 時点のモデルを用いた場合も同様であった．次元を減らすことで設定上は学習が易化したにも関わらず意味情報を再現出来なくなっていることから，画像刺激に反応しにくい脳領域にも意味情報の予測に必要な情報が含まれていたことが推測される．この結果は，Cukur ら [Cukur 13] が主張した，画像処理が脳の特定部位だけでなく広い領域にわたって行われているという説を示唆している．また，より高解像度（30,662 次元→ 65,665 次元）の全脳皮質情報を用いた方が全体的に性能が向上した [松尾 16] ということと併せて，ニューラルネットワークが高次元データのうち重要な要素を選び出す機能を持つことを示す例の一つとも言える．

6. おわりに

本稿では，深層学習モデル Encoder-Decoder Network による画像説明文生成システムを援用し，画像刺激に対する脳活動データと CNN により抽出される画像特徴量との対応関係を学習したモデルと組み合わせることで，深層学習を用いて脳活動データから人が想起している言語意味情報を説明文として出力する手法を提案した．学習に使用するモデルに関する 3 通りの実験設定に基づいて提案モデルを構築し，3 層のニューラルネットワークを用いたモデルにおいて最も生成文の精度が高くなるという結果を得るとともに，画像刺激を受ける脳活動データの自然言語文表現への変換を実現した．また，学習に用いる脳部位を限定して学習したモデルによる生成文を比較することで，人間の脳における画像認識処理が脳全体で行われていることを示唆する結果を得た．

今後の課題として，データの追加や数値設定の見直し，ベイズ最適化などの機械学習手法による各モデルの精度向上，脳活動→画像特徴量モデルにおける CNN の適用を検討している．

参考文献

[Bengio 08] Bengio,Y.,Lamblin,P.,Popovici,D.,Larochelle,H.:Greedy Layer-Wise Training of Deep Networks, In NIPS ‘06 (2006)

[Cho 15] Cho, K., Courville, A., Bengio, Y.: Describing Multimedia Content using Attention based Encoder Decoder Networks, Multimedia, IEEE Transactions on, 17(11): 1875-1886 (2015).

[Cukur 13] Cukur, T., Nishimoto, S., Hut, A. G., and Gallant, J. L.: Attention during natural vision warps semantic representation across the human brain, Nature Neuroscience 16 (2013).

[Huth 12] Huth, A. G., Nishimoto, S., Vu, A. T., Gallant, J. L.: A continuous semantic space describes the representation of thousands of object and action categories across the human brain, Neuron, 76(6):1210-1224 (2012).

[Huth 16] Huth, A. G., Lee, T., Nishimoto, S., Bilenko, N. Y., Vu, A. T., Gallant, J. L.: Decoding the semantic content of natural movies from human brain activity, Frontiers in Systems Neuroscience, 19(81) (2016)

[松尾 16] 松尾映里，小林一郎，西本伸志，西田知史，麻生英樹: 深層学習を用いた画像刺激による脳活動データからの説明文生成，人工知能学会第 30 回全国大会 (2016)

[Mikolov 13] Mikolov, T., Sutskever, I., Chen, K., Corrado, G., Dean, J.: Distributed Representations of Words and Phrases and their Compositionality, In NIPS’13 (2013).

[Mitchell 16] Mitchell,M.,Han,X.,Dodge,J.,Mensch,A.,Goyal,A.,Berg, A.,Yamaguchi,K.,Berg,T.,Stratos,K.,Daume III H.: Generating Natural Questions About an Image, In ACL’16 (2016)

[Nishida 15] Nishida, S., Huth, A. G., Gallant, J. L., Nishimoto, S.: Word statistics in large-scale texts explain the human cortical semantic representation of objects, actions, and impressions, Society for Neuroscience Annual Meeting 2015 333.13 (2015).

[Nishimoto 11] Nishimoto, S., Vu, A. T., Naselaris, T., Benjamini, Y., Yu, B., Gallant, J. L.: Reconstructing visual experiences from brain activity evoked by natural movies, Current Biology, 21(19):1641-1646 (2011).

[Simonyan 15] Simonyan, K., Zisserman, A.:Very deep convolutional networks for large-scale image recognition,In ICLR’15(2015).

[Sutskever 14] Sutskever, I., Vinyals, O., Le, Q. V.: Sequence to sequence learning with neural networks, In NIPS’14 (2014).

[Vinyals 15] Vinyals, O., Toshev,A., Bengio,S. and Erhan,D.: Show and tell: A neural image caption generator,InCVPR’15(2015).