1G4-OS-11a-5in2

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

Generating Descriptions for Brain Activity caused by Visual Stimulus with 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

The quantitative analysis of human brain activity based on language representation such as the semantic categories of words has been actively studied in the field of brain and neuroscience. Our study aims to generate natural language descriptions, instead of words, to human brain activation patterns caused by visual stimulus by employing deep learning method, which has gained more interest as an effective approach of automatically describing natural language expressions for various multi-modal information like images. We employ an image captioning system based on a deep learning framework as the basis for our method by learning the relationship between the brain activity data and the features of an intermediate layer of the deep neural network. We carried out three experiments and could generate natural language sentences which enable us to quantitatively interpret brain activity.

1. はじめに

近年，脳神経生理学の分野では，画像等の刺激を受けた際の脳活動パターンから人の想起する言語意味情報を解析する研究が盛んになっている．一方，自然言語処理の分野では，ニューラルネットワークを用いた深層学習（Deep Learning）の発展に伴い，画像に映る事象を言葉で説明する手法など数値で表される情報を自然言語文の形で表現する技術が開発されている．

これらの背景を踏まえ，本研究では，先行研究にて提案された画像説明文生成モデル [Vinyals 15, Xu 15] を脳活動データに適用し，脳活動の状態を解釈し自然言語文で説明する手法を実現することで，言語を介した脳活動の定量的理解を目指す．

2. 関連研究

2.1 言語表現による脳活動分析

近年，動画像などを視聴した際の脳の活動パターンから人がどのような意味カテゴリを想起しているかを調査する研究が盛んになってきており，多くの新しい知見が得られている [Mitchell 08, Nishimoto 11, Huth 12, Stansbury 13, Horikawa 13]．Huth ら [Huth 12] は，動画像中の物体や動作を類義語体系である WordNet[Miller 95] の語彙で表現し，脳神経活動との対応関係を捉えることで脳の皮質における言語意味のマップを作成した．Stansbury ら [Stansbury 13] は、潜在的意味解析手法 LDA[Blei 03] によるラベル付けを行うことで，静止画と語彙との対応関係，静止画と脳神経活動との対応関係を結びつけ，カテゴリに対する脳の意味解釈の活動領域を明確にするとともにモデルを構築した．このように，統計的言語モデルは脳活動における感覚や文脈の情報に基づく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．しかし，上述の先行研究においては脳活動と対応関係を学習する言語表現として単語意味カテゴリのみが対象となっている．本研究では，より記述力・説明力の高い自然言語文章を出力することで，脳活動の更なる定量的理解を目指す．

連絡先: 松尾映里，お茶の水女子大学理学部情報科学科小林研究室，〒 112-8610 東京都文京区大塚 2-1-1，g1220535@is.ocha.ac.jp

2.2 画像説明文の自動生成手法

入力画像を説明する自然言語文を出力する問題に対しては，主に 2 通りのアプローチがとられてきた．画像とその説明文からなるデータベースから入力画像に類似した画像を検索し，入力画像の説明文として類似画像に付けられた既存の説明文を再利用する手法 [Kuznetsova 12, Kuznetsova 14, Vendrov 16]，あるいは物体認識やシーン認識などにより特徴的な単語やフレーズを抽出し，位置関係や主述関係を解析しテンプレート文に当てはめる手法 [Elliott 13, Elliott 15, Kulkarni 13] である．しかし，これらのアプローチによって出力された画像説明文は文法的に正確ではあるものの機械的で柔軟性に欠け，表現力に限界があるという問題が指摘されている．それに対し，新奇な自然文を生成する手法が近年における深層学習の発展に伴って次々と開発されてきており，中でも機械翻訳 [Sutskever 14, Bahdanau 15] やメディア変換 [Chorowski 15] に用いられる Encoder-Decoder Network（Enc-DecNet）[Cho 14, Cho 15] に基づく研究が数多く報告されている [Donahue 15, Kiros 15, Mao 14, Vinyals 15]．Enc-DecNet は，Encoder と 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）して別の表現を出力する，深層学習モデルの枠組みである．Vinyals ら [Vinyals 15] は，Encoder として画像の特徴量抽出に効果的な GoogLeNet[Ioffe 15]（本手法では VGGNet[Simonyan 15]），Decoder として深層学習言語モデル LSTM-LM[Hochreiter 97, Sutskever 14] を採用した Enc-DecNet を構築することで，画像に対してその内容を説明する文の生成を実現した．また，Xu ら [Xu 15] は，同様の Enc-DecNet に Attention Mechanism[Bahdanau 15, Cho 15] を導入したモデルを提案し，生成文の精度向上を示した．Attention Mechanism は，Enc-DecNet に導入することで出力の各要素ごとに着目すべき入力要素を自動的に学習するシステムであり，画像の説明文を生成する手法においては，注目すべき画像の箇所を考慮した人間の情報処理に近いプロセスでの文生成を実現する．本手法では，上記画像説明文生成モデルを脳活動データに適用させ，画像刺激を受ける脳活動情報の言語化を行う Enc-DecNet の構築を試みる．

3. 提案手法

まず，先行研究 [Vinyals 15][Xu 15] における，深層学習を用いた画像説明文生成プロセスを説明する．

step 1. *Encoder*：*VGGNet* による特徴量の抽出
静止画を入力として VGGNet で画像特徴量を抽出．Attention Mechanism 適用時は VGGNet の途中層から 512 個の 14×14 次元データを，非適用時は VGGNet による処理を最後まで行った単一の 4096 次元データを Encoder の出力とする．

step 2. 中間表現の処理
Attention Mechanism 適用時は，step 1. において計算された中間表現集合の重み付き和を Decoder に渡す入力として算出．重み係数は 1 単語前の Decoder（LSTM）の隠れ状態と 512 個の中間表現から 3 層 MLP で計算される．非適用時は Encoder の出力をそのまま使用．

step 3. *Decoder*：*LSTM-LM* による単語予測
step 2. で計算された中間表現および 1 単語前の Decoder の隠れ状態を入力として，LSTM-LM で単語を出力．

step 4. 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 2-3 を繰り返し，1 語ずつ出力して文章を生成．

本提案手法は，上記の画像説明文生成プロセスを転用し，人の脳活動情報からその時見ている画像の内容を説明する文の生成を目指す．図 1,2 に概要図を示す．具体的には，画像刺激を受けているときの脳神経活動データと，VGGNet にその画像を入力して出力される特徴量，すなわち先行研究における中間表現との対応関係を学習したモデルを Encoder の代替とし，以降は同様の処理を行うことで先行研究モデルを利用し実現する．対応関係のモデルとして，3 層の多層パーセプトロン（Multi-Layer Perceptron：MLP）あるいは Ridge 回帰を用いている．

提案手法の処理の流れを以下に示す．

step 1'. 脳活動情報の中間表現への変換
同じ画像に対する脳活動データと VGGNet で計算される特徴量との対応関係を学習したモデル（3 層 MLP または Ridge 回帰）により，脳活動データを中間表現に変換する．

step 2〜4. 先行研究と同様の処理を行う．

4. 実験

本稿では，先行研究に基づく画像説明文生成モデルと中間表現と脳活動データとの対応関係の学習モデルについて，表 1 に示す 3 通りの組み合わせで脳活動データ説明文生成モデルを構築した．

表 1: 実験設定

	静止画→説明文	脳活動→中間表現
設定 1	Attention Mechanism	3 層 MLP
設定 2	Attention Mechanism	（ニューラルネット）
設定 3	非適用	Ridge 回帰

4.1 実験設定

システムの実装に際し，深層学習のフレームワーク Chainer[*1] 及び機械学習ライブラリ scikit-learn[*2] を利用している．

画像説明文生成モデルの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[*3] を使用する．本稿では，そのうち 168,000 データ分学習したモデルを実験に用いている．

∗1 http://chainer.org/
∗2 http://scikit-learn.org/
∗3 http://mscoco.org/

脳活動と画像中間表現の対応関係を学習するためのデータセットとして，動画像を被験者に見せた時の血中酸素濃度依存性信号（BOLD 信号;Blood Oxgenation Level Dependent Signal）を functional Magnetic Resonance Imaging（fMRI）を用いて記録した脳神経活動データ，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．脳活動データは 100×100×32 ボクセルのうち皮質に相当する 30,662 次元分のデータを扱い，実験 1 では 14×14×512=100,352 次元，実験 2，3 では 4,096 次元の中間表現との対応をとる．train 用データ数は 3,600 個であり，本稿では 20 回繰り返し学習させた．

4.2 実験 1：Attention 適用モデル＋ 3 層 MLP

まず，Attention Mechanism を用いた画像説明文生成モデルについて，4 に示すように，COCO の test 用画像からランダムに選んだ 2 つの脳活動データに対して説明文を生成し，学習が進み Attention が獲得されていることを確認した．

次に，脳活動データセットの test 用画像から選んだ 2 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を図 4 に示す．また，表 3 のように，画像説明文生成モデルについては train データ数毎の perplexity，3 層 MLP については train 周回毎の平均二乗誤差を記録し，その減少によって学習の進度を確認した．

出力された文は前置詞が無秩序に並んだ文になっており，画像の内容もあまり捉えられておらず，説明文として成立していない．平均二乗誤差の減少量も小さいことから，入力（30,662 次元）に対し出力（100,352 次元）が高次元すぎるために，MLP による脳活動データと中間表現集合との対応関係がうまく学習できなかったと推測される．

図 3: Attention Mechanism を用いた画像説明文生成．左から，生成元画像，生成文，主語出力時の Attention 可視化図

表 3: 実験 1：training 時の評価指標の変化

データ数	perplexity	周回数	平均二乗誤差
14000	88.67	1	118.32
42000	66.24	5	116.44
84000	60.40	10	114.31
126000	60.10	15	112.36
168000	60.32	16	112.01

4.3 実験 2：Attention 非適用モデル＋ 3 層 MLP

実験 1 と同様の手順で，Attention Mechanism を用いない画像説明文生成モデルについても，2 つの画像データに対して説明文が正しく生成されることを確認した．test 用画像から選んだ 2 つの脳活動データに対する説明文と画像を図 5 に，train データ数毎の画像説明文生成モデルの perplexity，train 周回毎の 3 層 MLP の平均二乗誤差を表 4 に示す．

Attention 適用時に比べ，意味のない前置詞がなくなり，出力語も画像の内容に近いものとなり，文法的にも意味的にもよ

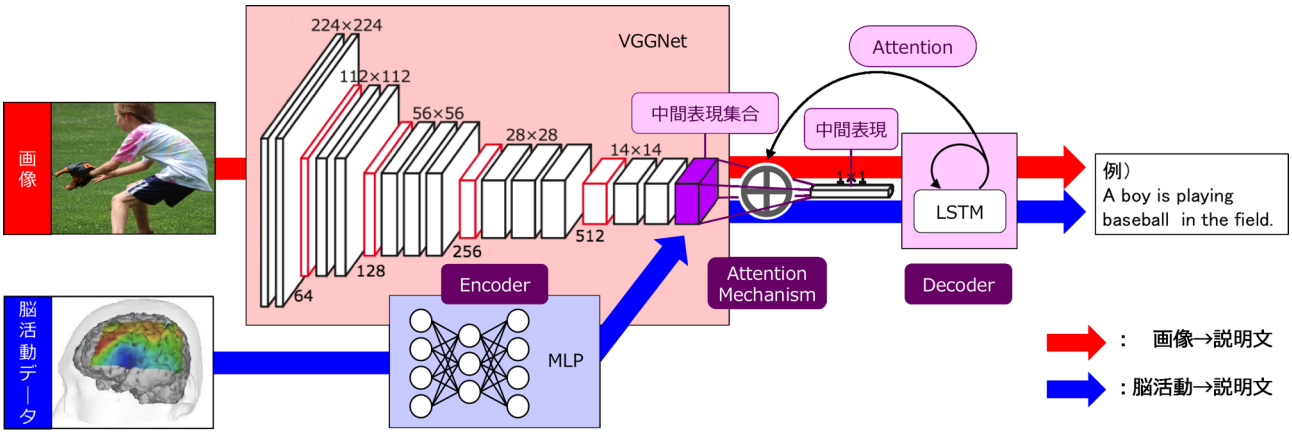

図 1: 本研究の概要図（実験 1：Attention Mechanism 適用モデル+ 3 層 MLP）

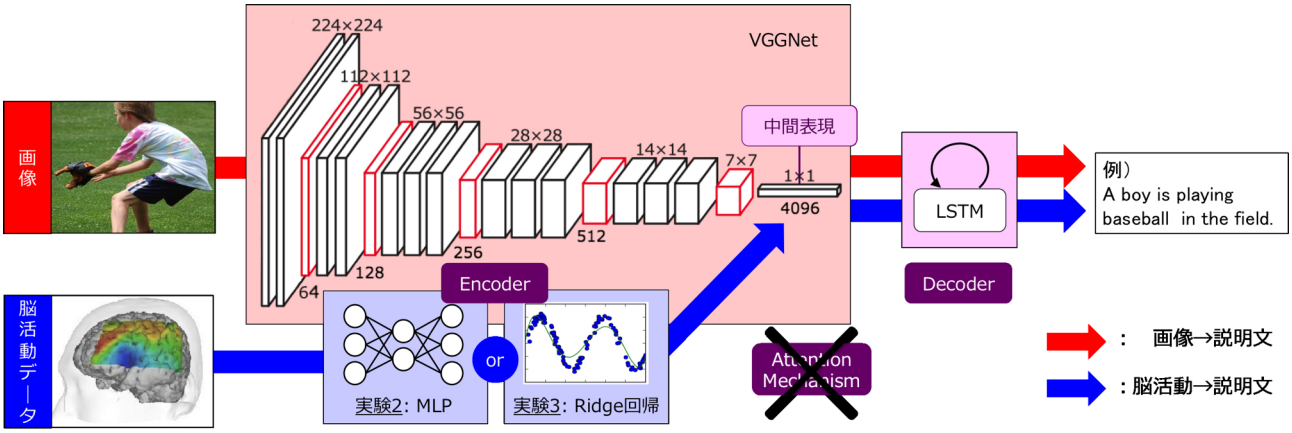

図 2: 本研究の概要図（実験 2：Attention Mechanism 非適用モデル+ 3 層 MLP or Ridge 回帰）

Table 2: 各パラメータ設定（詳細）

	静止画→説明文 1 Attention あり	静止画→説明文 2,3 Attention なし	脳活動→中間表現 1 3 層 MLP	脳活動→中間表現 2 3 層 MLP	脳活動→中間表現 3 Ridge 回帰
train データ	Microsoft COCO		動画刺激脳活動データ		
学習に関する ハイパーパラ メータ	学習率：1.0 (× 0.999) 勾配閾値：5 L2 正則化項：0.005		学習率：0.01 勾配閾値：5 L2 正則化項：0.005		L2 正則化項：0.5
学習する パラメータ	Attention および LSTM [-0.1,0.1] で初期化	LSTM のみ [-0.1,0.1] で初期化	3 層 MLP 重み係数 [-0.2,0.2] で初期化		Ridge 回帰パラメータ 0 で初期化
層ユニット数	各層 196	各層 1,000	30,662 - 1,000 - 100,352	30,662 - 1,000 - 4,096	
語彙	頻出語 3,469 語 (各 512 次元)		ー		
アルゴリズム	確率的勾配降下法		確率的勾配降下法		Ridge 回帰
誤差関数	交差エントロピー		平均二乗誤差		


A in of man people with and is on standing.


Man a people in of with and is on street.

図 4: 実験 1：生成した説明文とその時見ていた画像例


A group of people sitting next to each other.


Front of a white room with the street.

図 5: 実験 2：生成した説明文とその時見ていた画像例

り適切な表現を獲得している．平均二乗誤差の減少にも見られるように，中間表現の次元が 100,352 から 4,096 と大幅に低次元化したことで MLP の学習が順調に進んだと考えられる．

4.4 実験 3：Attention 非適用モデル+ Ridge 回帰

test 用画像から選んだ 2 つの脳活動データに対する説明文と画像を図 6 に示す．評価指標の値については，画像説明文生成モデルの perplexity は実験 2 と同様であり，Ridge 回帰の

表 4: 実験 2：training 時の評価指標の変化

データ数	perplexity	周回数	平均二乗誤差
14000	96.50	1	28.95
42000	47.87	5	22.70
84000	47.22	10	17.19
126000	47.37	15	13.37
168000	46.30	20	10.76

平均二乗誤差は 8.675 となった．
実験 1，2 と異なり，生成文は文章として完全に成立している．前置詞（in,on）や冠詞（an umbrella）の区別も正しくついており，文法が大幅に改善されていることがわかる．主語が「man」であることを除けば，画像の内容もかなり的確に捉えられている．特に 2 例目は「umbrella」を認識しながらも主語には人物を用いるなど，説明文として適切な出力が行われている．平均二乗誤差も実験 3 が最も低い値を示している．Attention 適用時よりも非適用時，非線形変換の 3 層 MLP 使用時よりも線形変換の Ridge 回帰使用時の方が精度が向上していることから，noisy なデータである脳活動に対しては，過学習を起こしにくい，より単純なモデルが適切なのではないかと推測できる．


A man is sitting on top of the table.


A man is in the back of an umbrella.

図 6: 実験 3：生成した説明文とその時見ていた画像例

5. おわりに

本稿では，画像刺激に対する脳活動データと VGGNet による画像特徴量との対応関係を学習し，深層学習モデル EncDecNet による画像説明文生成システムと組み合わせることで，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，Attention Mechanism を適用せず Ridge 回帰を用いた単純な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．

今後の課題として，train データの追加や数値設定の見直しによる精度向上，BLEU や METEOR などの指標を用いた実験結果の更なる評価および考察などが挙げられる．また，脳活動データと中間表現との対応関係を学習するモデルとして，3 層 MLP や Ridge 回帰に加えて新たに CNN の適用を予定している．白色化やベイズ最適化などの機械学習手法の採用も検討したい．

参考文献

[Bahdanau 15] Bahdanau, D., Cho, K., and Bengio, Y.: Neural machine translation by jointly learning to align and translate, In ICLR'15 (2015).

[Blei 03] Blei, D., Ng, A., Jordan, M.: Latent Dirichlet allocation, Journal of Machine Learning Research,3:993-1022(2003).

[Cho 14] Cho, K., Merrienboer, B., Gulcehre, C., Bougares, F., Schwenk, H., and Bengio, Y.: Learning phrase representations using RNN encoder-decoder for statistical machine translation, In EMNLP'14 (2014).

[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).

[Chorowski 15] Chorowski, J., Bahdanau, D., Serdyuk, D., Cho, K., and Bengio, Y.: Attention-based models for speech recognition, In arXiv preprint arXiv: 1506.07503 (2015).

[Donahue 15] Donahue, J., Hendricks, L.A., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., and Darrell, T.: Long-term Recurrent Convolutional Networks for Visual Recognition and Description, In CVPR'15 (2015).

[Elliott 13] Elliott, D., Keller, F.: Image description using visual dependency representations, In EMNLP'13 (2013).

[Elliott 15] Elliott D., Vries, A. P.: Describing Images using Inferred Visual Dependency Representations, In ACL'15 (2015).

[Hochreiter 97] Hochreiter, S., Schmidhuber, J.: Long Short-Term Memory, Neural Computation 9(8) (1997).

[Horikawa 13] Horikawa, T., Tamaki, M., Miyawaki, Y., Kamitani, Y.: Neural Decoding of Visual Imagery During Sleep, SCIENCE VOL 340 (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).

[Ioffe 15] Ioffe, S. and Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift, In arXiv preprint arXiv:1502.03167 (2015).

[Kiros 15] Kiros, R., Salakhutdinov, R., Zemel, R.: Unifying visual-semantic embeddings with multimodal neural language models, In NIPS'15 Deep Learning Workshop (2015).

[Kulkarni 13] Kulkarni, G., Premraj, V., Ordonez, V., Dhar, S., Li, S., Choi, Y., Berg, A. C., Berg, T. L.: Babytalk: Understanding and generating simple image descriptions, PA and MI, IEEE Transactions on, 35(12): 2891-2903 (2013).

[Kuznetsova 12] Kuznetsova, P., Ordonez, V., Berg, A. C., Berg, T. L., and Choi, Y.: Collective generation of natural image descriptions, In ACL'12 (2012).

[Kuznetsova 14] Kuznetsova, P., Ordonez, V., Berg, T. L., and Choi, Y.: TREETALK: Composition and compression of trees for image descriptions, In ACL'14 (2014).

[Mao 14] Mao J., Xu, W., Yang, Y., Wang, J., Huang, Z., Yuille, A.: Deep Captioning with Multimodal Recurrent Neural Networks (m-RNN), In ICLR'14 (2014).

[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).

[Miller 95] Miller, G. A.: WordNet: a lexical database for English, Communications of the ACM, Volume 138,Pages 39-41(1995).

[Mitchell 08] Mitchell, T. M., Shinkareva, S. V., Carlson, A., Chang, K., Malave, V. L., Mason, R. A., Just, M. A.: Predicting Human Brain Activity Associated with the Meanings of Nouns, Science 320, 1191 (2008).

[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).

[Stansbury 13] Stansbury, D. E., Naselaris, T., Gallant, J. L.: Natural Scene Statistics Account for the Representation of Scene Categories in Human Visual Cortex, Neuron 79, pp.1025-1034, September 4, 2013, Elsevier Inc (2013).

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

[Vendrov 16] Vendrov, I., Kiros, R., Fidler, S., Urtasun, R.: Order-Embeddings of Images and Language, In ICLR'16 (2016).

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

[Xu 15] Xu, K., Ba, J., Kiros, R., Courville, A., Salakhutdinov, R., Zemel, R., and Bengio, Y.:Show, attend and tell: Neural image caption generation with visual attention, In ICML'15(2015).