The 33rd Annual Conference of the Japanese Society for Artificial Intelligence, 2019

深層学習を用いた動画刺激時の脳活動データからの文生成

Generating Natural Language Descriptions with Brain Activity Data Evoked by Video Stimuli using Deep Learning

張 嘉瑩 *1 Kaei Cho　西田 知史 *2 Satoshi Nishida　西本 伸志 *2 Shinji Nishimoto　小林 一郎 *1 Ichiro Kobayashi

*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

Quantitative analyses 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 video stimuli by employing deep learning. Due to the lack of brain training data, the proposed method employs a pre-trained S2VT (end-to-end sequence-to-sequence model to generate captions for videos). To apply brain activity data to the video captioning model, we train a model to learn the corresponding relationship between brain activity data and video features. As result of experiments, we have not yet been successful in generating appropriate sentences. We will further devise the architecture.

1. はじめに

近年，脳神経科学分野において，脳神経活動の意味表象情報を定量的に理解する研究が盛んに行われている．なかでも，近年の深層学習の成果を取り入れて脳活動データを解読する研究が増えている．本研究では，人が動画刺激によって頭の中で知覚している意味情報，すなわち動画によって知覚された事象を，functional Magnetic Resonance Imaging (fMRI) を用いて観測された脳活動データによって，自然言語文で説明する深層学習手法を提案する．一般に，fMRI を用いた脳活動データの収集コストが大きく，また脳のサイズに個人差があるために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 video captioning 手法を援用することで少量データを効率的に活用する手法を採用する．

2. 関連研究

脳活動データを用いて人が知覚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報告されている [Nishimoto 11, Huth 12, Cukur 13, 松尾 17] ．Cukur ら [Cukur 13]，Huth ら [Huth 12] らは，動画像中の物体に注目し，ラベル分類に基づき単語レベルの意味表象推定を対象とした分析を行った．松尾ら [松尾 17] は，動画像刺激時における脳神経活動から，深層学習の手法を用いてより説明力の高い自然言語文によって意味情報を表現した．その際，image captioning 手法 [Vinyals 15] を援用することによって，観測コストが高く大規模なデータ収集が困難な fMRI データの効果的活用を行なった．本研究では，松尾ら [松尾 17] による少量データの効果的活用手法を参考にし，時空間情報を含む動画を視聴した際の脳神経活動から動画を説明する自然言語文を出力することで，脳神経活動の更なる定量的な理解を目指す．

連絡先: 張嘉瑩，お茶の水女子大学院　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，g1420526@is.ocha.ac.jp

3. 動画刺激からの説明文生成

本提案手法では 3.1 節と 3.2 節で示す 2 種類のモデルを組み合わせることにより，脳活動データを入力し，その人が知覚している意味情報を説明する自然言語文の生成を目指す．図 1 に提案手法の概要を示す．

3.1 モデル① : 動画→特徴量→説明文

動画から自然言語文を生成する video captioning 手法として，Venugopalan ら [Venugopalan 15] によって提案されたモデル (S2VT) を使用した．S2VT は sequence-to-sequence モデル [Sutskever 14] を使用しており，stacked LSTM によってそれを可能としている．stacked LSTM では，まず Convolutional Neural Network (CNN) によるフレームごとの RGB 画像もしくは optical flow [Brox 2004] の特徴量を変換 (encode) し，全てのフレームが読み込まれた後に一語ずつ復号 (decode) して文生成をする．Encoder では，動画ごとにフレームの特徴量を time step で入力することで，動画から自然言語文を生成することを可能にしている．

今回は，CNN に VGGNet [Simonyan 15]，その CNN の入力として RGB 画像，また 2 層の stacked LSTM を使用し，動画→特徴量→説明文モデルとした．

3.2 モデル② : 脳活動データ→特徴量

S2VT を脳活動データに適用するために，脳活動データを入力として，被験者が見ているとされる動画フレームより抽出される特徴量を予測するモデル，つまり脳活動データを入力として S2VT の sequence-to-sequence モデルにおける入力を予測する予測モデルに 3 層 Neural Network(NN) を使用した．

3.3 処理の流れ

以下に提案手法における処理の流れを示す．

step 1-1. 動画から特徴量への変換

モデル①における前処理として，動画を frame ごとに clipping し，VGGNet で特徴量に変換する．

step 1-2. 特徴量から文生成の学習

step 1-1. で得られた特徴量を，動画ごとに Encoder に time step で入力し，Decoder では一語ずつ単語を出力することによってモデル①を学習する．

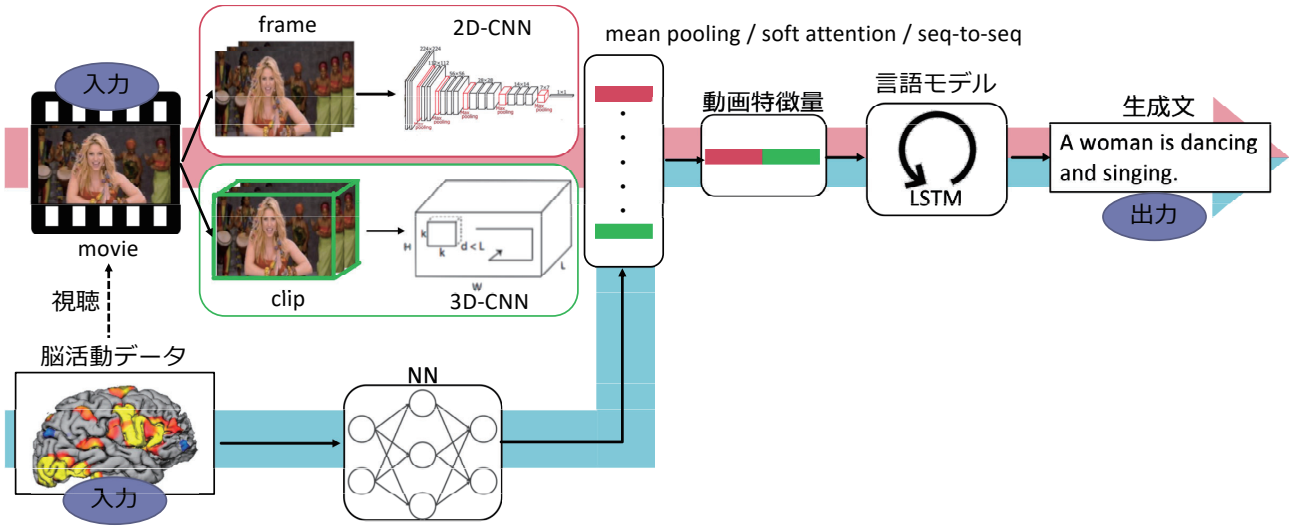

図 1: 本研究の概要

step 2. 脳活動データから特徴量の予測

脳活動データと動画 frame ごとの特徴量 (step 1-1. で変換後のもの) の対応関係を学習したモデル②を用いて，脳活動データから特徴量を予測する．

step 3. 脳活動データによる特徴量から文生成

step 1-2. で学習済みのモデルモデル①に，step 2. で予測された特徴量を入力することにより，文生成をする．

4. 実験

4.1 実験 1：動画→特徴量→文生成モデル

4.1.1 実験設定

学習のためのデータセットとして 1,970 の動画とその説明文ペアからなる Microsoft Video description corpus(MSVD) [Chen 11] を使用する．このデータセットは Youtube clips から短く，かつ単独の動作のみが含まれた clip を選んだものであり，動画の長さは平均 10.2 秒である．実装には，動画から特徴量の変換では深層学習フレームワークの Caffe を，特徴量から文生成の学習では TensorFlow を用いたコード [*1] を使用した．動画の前処理として各動画を 1 frame ごとに clipping し，80 frames より多いものに関しては線形に等間隔に 80 frames に直した．これは，学習で使用している LSTM が 80 time steps で固定されているためである．また，80 frames に満たない動画に対しては 0 を padding した．学習に関する詳細設定は表 1 の左列に示す．

4.1.2 実験結果と考察

epoch 毎に test データの loss を記録し，その値の減少により学習の進度を確認した．また，test データの動画入力による文生成の結果例を表 2 に示す．1 例目は十分に妥当な説明文が生成され，2 例目は目的語が違うものも主語と述語は正確に捉えられている．また，文章全体に大きな文法の間違いなどはなく，動画内容を大まかに捉えた可読性のある自然言語文の生成ができたと言える．

4.2 実験 2：脳活動データ→特徴量モデル

4.2.1 実験設定

データセットとして，動画を被験者に視聴させた時の血中酸素飽和度信号 (BOLD 信号) を fMRI を用いて 1 秒ごとに記録した脳活動データ，およびその動画を使用する．刺激として用いた動画は，映画，自然，アニメ，機械など様々な種類のものを含んだ十数秒の動画となっている．また脳活動データは，脳活動の観測領域 96×96×72 ボクセルのうち，皮質に相当する 62,552 次元のデータを使用し，1 秒ごとに clipping した動画による特徴量との対応関係を学習する．このとき，動画を視聴してからその反応が脳活動データに現れるまでの遅延は一般に 3 6 秒とされていることから，今回は 3 秒と仮定して実験を行なった．実装には，深層学習フレームワークの Chainer を用いた．学習に関する詳細設定は表 1 の右列に示す．

4.2.2 実験結果

epoch 毎に test データの平均二乗誤差を記録し，収束することを確認した．

4.3 実験 3：特徴量→文生成モデル

4.3.1 実験設定

実験 1，実験 2 で学習した動画→特徴量→説明文モデルと脳活動データ→特徴量モデルを組み合わせることにより，脳活動データを入力とした際の説明文生成を実行した．また，その時見ている動画から直接モデル①を使用した説明文生成も行なった．この動画は 476clips あり，実験 2 で学習した際の train データと test データは，351clips と 125clips に相当する．

4.3.2 実験結果と考察

脳活動データを入力とした際の説明文生成，および動画から直接モデル①を使用した説明文生成のいずれも全ての出力文がほとんど同じもの (例:A person is cleaning the floor.) になった．原因としては，モデル①を学習した際には encoder の入力が 80 time step であったが，実験 2 で変換した特徴量は多くても 20 frames 分しか入力で使えないために，長さ 80 に対してほとんどを 0 で padding しており，情報が削減されていることが考えられる．

*1 https://github.com/chenxinpeng/S2VT

表 1: 詳細学習設定

	①動画→特徴量→説明文	②脳活動データ→特徴量
データセット	MSVD	動画刺激による脳活動データ
データ数 (train/test)	1,576 / 394	6,000 / 1,200
アルゴリズム	Adam	SGD
学習に関する ハイパーパラメータ	encoder step：80 decoder step：20 学習率：0.0001 epoch：1000	学習率：0.01 勾配閾値：1 L2 正則化項: 0.003 epoch：100
層ユニット数	各層 1000	62,552 - 6,000 - 4,096
誤差関数	交差エントロピー	平均二乗誤差

表 2: 動画から生成した説明文の例

A hamster is eating.

A man is slicing a potato.

5. おわりに

本研究では，sequence-to-sequence による video captioning 手法 (S2VT モデル) を援用し，動画刺激に対する脳活動データと CNN によって抽出される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，脳活動データから人が知覚している言語意味情報を自然言語文として出力する手法を提案した．文章は生成されたが，全ての出力文がほとんど同じになっており，動画の時空間情報を加味した生成文が出力されたとは言い難い．

今後の課題として，モデル①の改良として MSVD の約 5 倍のデータ数を持つ MSR-VTT をデータセットとして使用することや，attention を用いることが考えられる．また，実験 3 においてモデル①の encoder の入力が 80 time step であることから，1 clip からより多くの特徴量を出力できるよう，モデル②についても検討を重ねたい．

謝辞

本研究の一部は，科研費新学術領域研究 (課題：18H05118) の支援を受けたものである．

参考文献

[Brox 2004] T. Brox, A. Bruhn, N. Papenberg, and J. Weichkert: High accuracy optical flow estimation based on a theory for warping, ECCV, pages 25-36 (2004).

[Chen 11] D. L. Chen, and W. B. Dolan: Collecting highly parallel data for paraphrase evaluation, in ACL (2011).

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

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

[松尾 17] 松尾映里, 小林一郎, 西本伸志, 西田知史, 麻生英樹: 画像説明文生成手法を援用した画像刺激時の脳活動の説明文生成, 言語処理学会, P6-2 (2017).

[Ngiam 11] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng: Multimodal deep learning, In Intl. Conf. on Machine Learning (2011).

[Nishimoto 11] S. Nishimoto, A. T. Vu, T. Naselaris, Y. Benjamini, B. Yu, and J. L. Gallant.: 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).

[Venugopalan 15] S. Venugopalan, M. Rohrbach, J. Donahue, T. Darrell, R. Mooney, K. Saenko: Sequence to sequence - video to text, The IEEE International Conference on Computer Vision (ICCV) (2015).

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