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

松尾映里† 小林一郎† 西本伸志‡ 西田知史‡ 麻生英樹¶
†お茶の水女子大学 ‡情報通信研究機構 ¶産業技術総合研究所
†{g1220535,koba}@is.ocha.ac.jp, ‡{nishimoto,s-nishida}@nict.go.jp,
¶h.asoh@aist.go.jp

1 はじめに

近年，脳神経生理学の分野では，脳の活動パターンから人の想起している言語意味情報を解析する研究が盛んになっている．本研究は，fMRI で観測した画像視聴時の脳活動データから，人が画像刺激によって頭の中に抱いた意味表象，すなわち画像によって想起された事象を，深層学習を用いて自然言語文で説明する手法を構築する．しかし，fMRI により観測する脳活動データは取得のためのコストが大きく，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，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う．

2 関連研究

画像刺激を受けた際の脳活動データから人が想起している言語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の意味の対応関係を捉えることで実現されている．Huth ら [1] は、動画像中の物体や動作を類義語体系である WordNet 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 [2] は、潜在的意味解析手法 LDA（Latent Dirichlet Allocation）によるラベル付けを行い，静止画と語彙，および静止画と脳神経活動との対応関係を結びつけるモデルを構築した．しかし、これらの先行研究は単語のみによる意味表象の推定を対象としている．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．

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，脳活動データを入力としてそのとき人が想起している内容を説明する自然言語文の生成を目指す．図 1 に概要図を示す．

3.1 画像→説明文モデル

本手法の基盤として，機械翻訳やメディア変換によく用いられる深層学習のモデルである Encoder-Decoder Network[3] を用いて実装される，画像に映る事象を自然言語で説明するキャプション付けのモデルを用いる [4]．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[5] を Encoder，2 層 LSTM-LM[3] を Decoder とした画像→画像特徴量→説明文モデルを採用する．学習には画像とその説明文を用い，脳活動データは扱わない．

3.2 脳活動データ→画像特徴量モデル

画像刺激を受ける被験者の脳活動情報を入力とし，そのとき見ている画像から VGGNet によって抽出される画像特徴量を予測，すなわち脳活動データを上記の画像→説明文モデルにおける中間表現に変換するモデルを用いる．適切な手法を探るため，Ridge 回帰，3 層 Neural Network（NN），5 層 Deep NN（DNN）の 3 通りの実装を行い比較した．また，少量データの効果的利活用のため，DNN については積層 AutoEncoder[6] による事前学習手法を適用した．学習には脳活動データとその時見ている画像を用い，文章は扱わない．

4 処理の流れ

実行時の処理は以下のようになる．

step 1. 脳活動情報の中間表現への変換
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．これを中間表現とし，以降の Step では画像→説明文モデルの処理を行う．

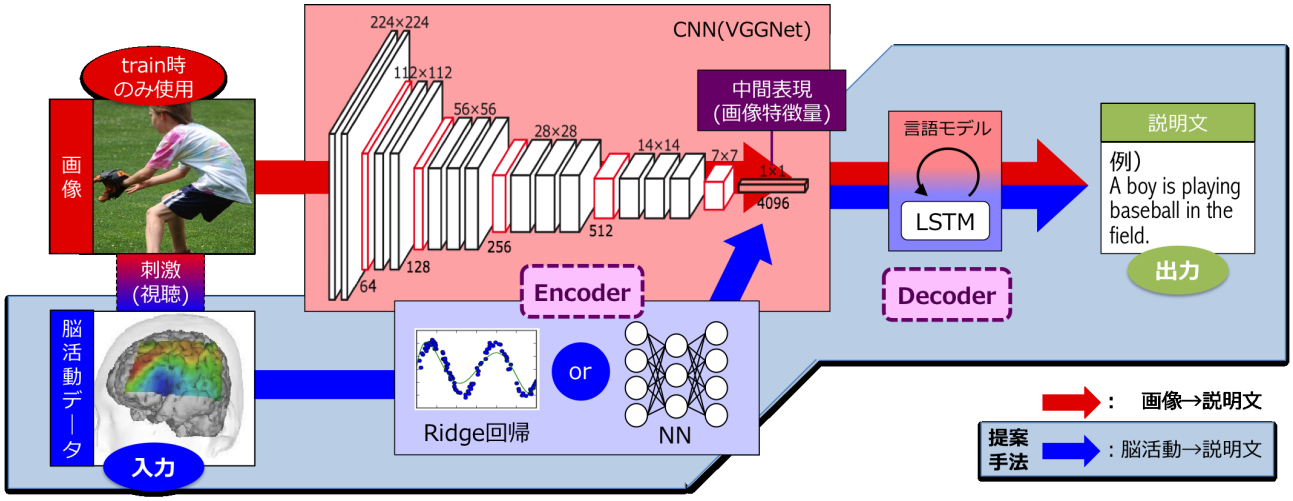

図 1: 本研究の概要図

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．

このように画像→画像特徴量→説明文モデル，脳活動データ→画像特徴量モデルを各々事前に学習させ，実行時に組み合わせることで脳活動データ→画像特徴量→説明文モデルを実現する．

5 実験

システムの実装に際しては，深層学習のフレームワーク Chainer[1] を利用した．

5.1 実験 1：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．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 windowsize=5 で事前学習した word2vec 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．また，train データ中に 50 回以上出現した 3,469 語を説明文生成に使用する語彙とした．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

図 2 のように epoch 毎に test 用画像からの出力文の perplexity を記録し，その減少により学習の進度を確認した．また，test 用画像からランダムに抽出した 2 つの画像に対して生成した説明文を図 3 に示す．

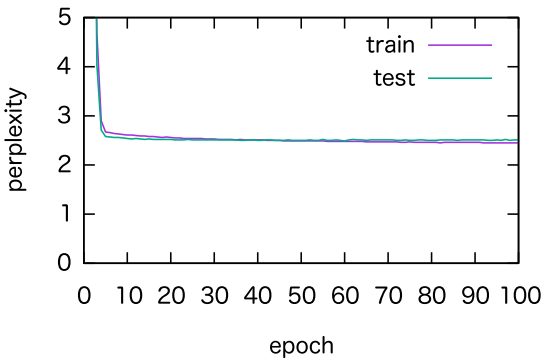

図 2: 実験 1 training 時の評価指標の変化

A man is surfing in the ocean on his surfboard.

A black and white cat is sitting on the toilet.

図 3: 画像から生成した説明文の例，画像はランダムに抽出

[1] http://chainer.org/
[2] http://mscoco.org/

表 1: 各パラメータ設定（詳細）

	画像→画像特徴量→説明文モデル	脳活動データ→画像特徴量モデル		
		1: Ridge 回帰	2: 3 層 NN	3: 5 層 DNN
train データ	Microsoft COCO	動画刺激による脳活動データ		
学習量	414,113sample×100epoch	4,500sample×1,000epoch		
アルゴリズム	Adam	Ridge 回帰	確率的勾配降下法	
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		

5.1.3 考察

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，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 2: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，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．立体撮像 96×96×72 ボクセル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，その時見ている静止画から 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，教師なし脳活動データ 7,540sample を用いて AutoEncoder による事前学習を各層 200epoch 行い，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

図 4 のように epoch 毎に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．

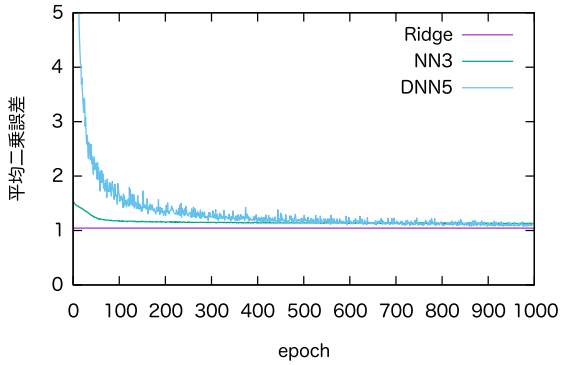

図 4: 実験 2 training 時の評価指標の変化

5.2.3 考察

平均二乗誤差において 3 種モデルはいずれも収束した．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，そのままでは解釈困難な画像特徴量では評価せず，次節の実験 3 にて 3 種モデルの比較・考察を行う．

5.3 実験 3：脳活動データ→説明文モデル

実験 1 で学習した画像→説明文モデルと実験 2 で学習した脳活動→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画像→説明文モデルを用いた説明文生成も行った．

	Ridge	3層NN	5層DNN	画像→説明文モデル
	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.

図 5: 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

5.3.1 実験結果

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説明文およびその時の画像と，画像→説明文モデルによる説明文を図 5 に示す．

5.3.2 考察

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た．実験 1 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，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 2 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，Ridge 回帰，3 層 NN を用いたモデルについてはおおむね画像の内容を捉えた類似の説明文が生成されたが，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，また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．興味深いのは，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．

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．

今後の課題として，データの追加や数値設定の見直し，白色化やベイズ最適化などの機械学習手法の採用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 の適用や，使用する脳活動の視覚野など特定の部位への限定を検討している．また，BLEU や METEOR などの指標を用いた実験結果の更なる客観的評価・分析も実施したい．

参考文献

[1] A. G. Huth, S. Nishimoto, A. T. Vu, J. L. Gallant, ”A Continuous Semantic Space Describes the Representation of Thousands of Object and Action Categories across the Human Brain,” Neuron, 76(6):1210-24, 2012

[2] D. E. Stansbury, T. Naselaris, J. L. Gallant,”Natural Scene Statistics Account for the Representation of Scene Categories in Human Visual Cortex,” Neuron, 79(5):1025-34, 2013

[3] K. Cho, A. Courville, Y. Bengio. ”Describing Multimedia Content using Attention-based Encoder-Decoder Networks.” CoRR, abs/1507.01053, 2015.

[4] O.Vinyals, A.Toshev, S.Bengio, D.Erhan, ”Show and tell: a neural image caption generator,” in CVPR’2015, 2015.

[5] K. Simonyan, A. Zisserman,”Very deep convolutional networks for large-scale image recognition,”in ICLR, 2015.

[6] Y. Bengio, P. Lamblin, D. Popovici and H. Larochelle, ”Greedy Layer-Wise Training of Deep Networks,” in NIPS‘2006, 2006.