深層学習による画像説明文生成手法の脳活動データへの適用

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，画像等の刺激を受けた際の脳の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を用いて表現する技術が開発されている．

これらの背景を踏まえて，本研究では，Xu らによって提案された画像説明文生成モデル [1] を脳神経活動データに適用し，脳活動の状態を解釈し，記述力の高い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

画像刺激を受けた際の脳活動データを入力としてその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 [2] 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 [3] 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るモデルを構築した．しかし、これらの先行研究は単語の推定のみを対象としており、より記述力・説明力の高い自然言語文の生成による脳活動データの解釈を行う研究は行われていない．

一方で，深層学習を用いて画像に映る事象を言葉で説明するキャプション付けの研究は、既に多くの先行研究が報告されている [1][4]．その中でも，本研究では画像の特定の箇所など，入力情報において着目すべき情報を捉え性能向上をもたらす Attention Mechanism を導入した Xu ら [1] の提案モデルを基礎としたモデルを構築し，脳活動データとの対応をとることで人の思考内容を説明する文を生成する手法への転用を目指す．

3 Encoder-Decoder Network

Encoder-Decoder Network（Enc-DecNet）とは，機械翻訳やメディア変換に用いられる深層学習のモデルである [5]．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という形で実現される．

先行研究 [1] では，Encoder に VGGNet[6]，Decoder に LSTM-LM[5] を採用した Enc-DecNet に Attention Mechanism を導入したモデルを構築している．VGGNet の出力した中間表現と脳活動データとの対応関係を多層パーセプトロン（Multi-Layer Perceptron；MLP）を用いて学習する本提案モデルも，Encoder として MLP を導入した Enc-DecNet と見なしうる．

3.1 CNN（VGGNet）

先行研究 [1] で Encoder として用いる VGGNet は，画像の特徴量抽出に効果的な深層学習のモデルである Convolutional Neural Network（CNN）の一種である [6]．

CNN は脳の視覚野における神経科学の知見を基に開発されたモデルであり，多チャネルの画像に小サイズの二次元フィルタを畳み込む演算を行うことで画像の持つ局所的な特徴を抽出する Convolution 層と，その多チャネル画像の小領域での値を一つの値に集約し解像度を落とすことで抽出された特徴の位置が若干変化しても取り出される特徴はほとんど変化しないという特徴の不変性を獲得する Pooling 層を複数積み重ね，最後に通常の全結合層を数層重ねて出力を計算する．

3.2 RNN-LM（LSTM-LM）

Decoder として用いる Long Short-Term Memory-Language Model（LSTM-LM）は，時系列データ対応

の深層学習モデル Recurrent Neural Network（RNN）による言語モデル RNN-LM の一種である [5].

RNN は隠れ状態（計算時の変数）の情報を次時刻の入力とすることで過去の履歴を利用した時系列解析を行うモデルであり，RNN-LM は過去の文脈（t-1 個の単語）から t 番目の単語として各語が選ばれる確率を算出する．1 時刻前の隠れ状態（時刻 1〜t-1 の単語情報），1 時刻前の予測結果（時刻 t-1 の単語），外部情報（本稿では中間表現に相当）の 3 つを入力とし，逐次的に次の単語の予測を繰り返して文章を生成する．

3.3 Attention Mechanism

Attention Mechanism[5] 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る．画像の説明文を生成する手法においては，各語生成時に画像のどこに注目すべきか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．

従来の Enc-DecNet では Encoder の出力した単一の中間表現をそのまま Decoder の入力として与えるが，Attention Mechanism では，Encoder に複数の中間表現を出力させ，各中間表現に重み係数（注目度）をかけた重み付き和を Decoder の入力として与える．重み係数は各時刻ごとに 1 時刻前の Decoder の状態と中間表現を入力とした 3 層 MLP で計算され，深層学習のモデルの一部として同時に学習される．

4 提案手法

まず，先行研究 [1] 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 で特徴量を抽出．Attention Mechanism 適用のため VGGNet の途中の Pooloing 層で処理を打ち切り，全結合層直前の 512×14×14 次元のものを Encoder の出力とする．出力された中間表現集合は静止画を重複ありで 512 個に分割した 14×14 小領域の特徴量に相当する．

step 2. Attention Mechanism による重み付き和処理
step 1. において計算された中間表現の集合に対し，1 時刻前の Decoder（LSTM）の隠れ状態を元に MLP で学習した重み係数をかけ，重み付き和を導出．

step 3. Decoder；LSTM-LM による単語予測
step 2. において計算された重み付き和，および 1 時刻前の Decoder（LSTM）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することで，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章の生成を目指す．図 1 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，脳の視覚神経の働きを基に構成されたモデルであり脳活動データとの相関関係が期待できる VGGNet にその画像を入力して出力される画像小領域の特徴量，すなわち先行研究における中間表現集合との対応関係を 3 層 MLP で学習して Encoder の代替とし，それ以降は同様の処理を行うことで先行研究の学習結果を利用し実現する．

提案手法の処理の流れを以下に示す．

step 1'. MLP による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 の出力との対応関係を学習した 3 層 MLP により，脳活動データから中間表現を算出する．

step 2〜4. 先行研究と同様の処理を行う．

5 実験

5.1 画像に基づく文生成

本研究では，まず Xu ら [1] の画像説明文生成モデルを構築し，その有効性を確認するとともに，ハイパーパラメータ値設定による学習結果の変化を観察した．

5.1.1 実験設定

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，train，test 用データセットとして静止画とその説明文のペアからなる Microsoft COCO[2] を使用した．本研究では 414,113 個の train 用データのうち，94,500 個まで学習した結果を提示する．

学習に関するハイパーパラメータの数値設定については，学習率を 0.001 とする先行研究と同様の設定と，Chainer で採用されている深層学習の効率化手法を取り入れ，学習率を 1.0（パラメータ更新毎に ×0.999），勾配閾値 5，L2 正則化項 0.005 とした設定の，2 通りについて実験を行った．その他のハイパーパラメータは VGGNet の出力次元に揃え，各語は 512 次元ベクトルで表現し，LSTM のユニット数は各層 14×14=196 に設定した．また，train 用データ中に 50 回以上出現した 3,469 語を説明文生成に使われる語彙とした．

学習するパラメータは Attention Mechanism および Decoder（LSTM）の重み係数とし，[-0.1,0.1] でランダムに初期化した．Encoder（VGGNet）は事前学習したものを用い，更新を行わない．学習アルゴリズムは確率的勾配降下法，誤差関数は交差エントロピーを使用．

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

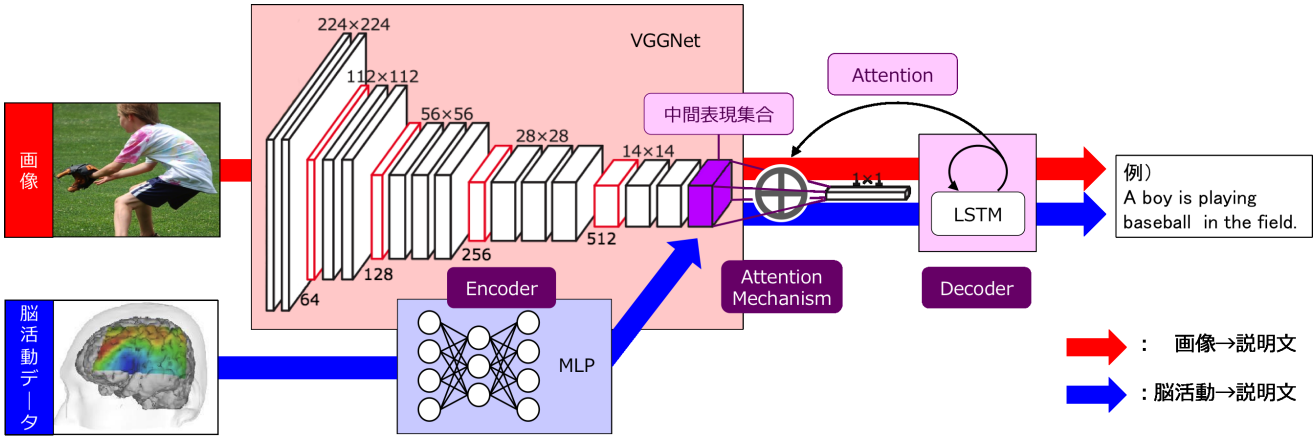

図 1: 本研究の概要図.

5.1.2 実験結果

設定した 2 通りのハイパーパラメータ（先行研究／効率化）について，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，およびその主語生成時の Attention の重みを，それぞれ図 2，図 3 に示す．また，表 1 のように train データ数毎に出力文の perplexity を記録し，その減少により学習の進度を確認した．

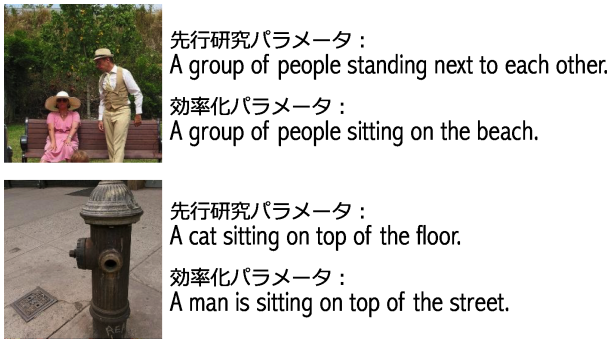

図 2: 生成した説明文の例，画像はランダムに抽出

図 3: 主語生成時の Attention を白く可視化した例

表 1: training 時の perplexity の変化

データ数	先行研究	効率化
7000	147.83	240.17
24500	66.52	69.47
42000	50.87	66.24
59500	42.96	79.74
77000	37.77	64.35
94500	35.04	61.59

5.1.3 考察

出力された説明文は逐次出力の文章としては文意を読み取ることが十分可能であり，画像を正確に説明できていない要素も見受けられるものの，おおむね画像の大意を認識し表現していると評価できる．興味深いのは，どちらのハイパーパラメータ値設定でも人間は認識できているが，ポストは cat あるいは man と誤認している点である．これは，使用した train データに人物画像が多く含まれるのに対し，ポストの画像は数個しか存在しないことから，ポストという概念の獲得にはデータ量が不十分であったことが原因の一つと考えられる．

ハイパーパラメータ値の設定による差が顕著に現れたのは，Attention の学習結果である．先行研究の設定では Attention の学習が不十分だが，効率化手法による設定では画像中の注目すべき部分を的確に捉えており，導入手法が深層学習の学習効率を向上させたと推測される．一方，生成文および perplexity は先行研究の方が優れており，今後学習が進んで Attention が獲得されれば，値が適切に調整されている先行研究の方が全体として良い結果となる可能性が考えられる．

5.2 脳活動データに基づく文生成

上記画像説明文生成モデルを基に，脳活動データを入力としてその時見ている画像の説明文を出力するシステムを構築した．

5.2.1 実験設定

先行研究と同様，深層学習のフレームワーク Chainer を利用した．train，test 用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）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（静止画像）を使用する．脳活動データは 100（x）×100（y）×32（z）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，512×14×14=100,352 次元の中間表現との対応関係を 3 層 MLP で学習する．画像のサイズは VGGNet の入力次元に揃え 224×224 とし，train 用データ数は 3,600（2 秒毎に 7,200 秒分記録）である．

学習を行う MLP のハイパーパラメータ値設定については，学習率 0.01（パラメータ更新毎に ×0.999），勾配閾値 5，L2 正則化項 0.005，中間層ユニット数 1000 に設定した．学習するパラメータは [-0.1,0.1] でランダムに初期化し，学習アルゴリズムは確率的勾配降下法，誤差関数は平均二乗誤差を用いている．

基となる画像説明文生成モデルには，効率化手法を導入したパラメータ値設定のものを使用している．

5.2.2 実験結果

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 2 のように train 周回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．

A man in front on the street room at tennis.

Field in the a man bathroom room people and kitchen.

図 4: 生成した説明文およびその時見ていた画像例

表 2: training 時の平均二乗誤差の変化

周回数	平均二乗誤差
1	29346.12
3	9902.90
5	9092.57
7	9038.79
9	9038.41

5.2.3 考察

出力された説明文は文章として成立しておらず，画像の意味内容もあまり捉えられていない．また平均二乗誤差も周回数に比して減少量が小さい．これは，入力（30,662 次元）に対し出力の次元（100,352 次元）が大きい，ハイパーパラメータ値の設定が不適切，あるいは train 周回数および train データ数の不足などの理由で，MLP による脳活動データと VGGNet の Pooling 層との対応関係の学習がうまくいかなかったことが原因であると推測される．中でも、出力次元数の大きさは特に学習を困難にしていると考えられる．また，画像からの説明文生成時にはどのような入力に対しても主述関係などの自然言語らしい構文は保たれていたにも関わらず，モデルの転用によりその知識が失われてしまっている点も考察の余地が残る．

6 おわりに

本稿では，深層学習モデル Enc-DecNet に Attention Mechanism を導入した画像説明文生成システムを構築し，その有効性を確認した．また，MLP を用いて脳活動データと CNN の pooling 層との対応関係を学習し，構築したシステムを転用す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が．提案手法に改善の余地があることが確認された．

今後の課題として，動画説明文生成手法については 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評価および考察，他手法との比較などが挙げられる．脳活動説明文生成手法については，学習対象となる CNN の出力層の低次元化が挙げられる．また，ベイズ最適化を採用した最適パラメータの発見なども検討したい．

参考文献

[1] K. Xu, J. L. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. S. Zemel, and Y. Bengio, ”Show, attend and tell: Neural image caption generation with visual attention,” in ICML ’2015, 2015.

[2] 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

[3] 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

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

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

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