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　(指導教員：小林 一郎)

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，先行研究にて提案された画像説明文生成モデル [1][2] を脳神経活動データに適用し，脳活動の状態を解釈して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　深層学習を用いた文生成

本研究では，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 Encoder-Decoder Network（Enc-DecNet）を用いる [3]．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．

Vinyals ら [1] は，Encoder として画像の特徴量抽出に効果的な GoogLeNet（本手法では VGGNet[4]），Decoder として深層学習言語モデル LSTM-LM[3] 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 [2] は，同様の Enc-DecNet に Attention Mechanism[3] 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．

3　提案手法

まず，先行研究 [1][2] 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 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 3 層の多層パーセプトロン（Multi-Layer Perceptron：MLP）で学習して 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．

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．

4　実験

Xu ら [2] による Attention Mechanism 適用モデル／Vinyals ら [1] による非適用モデルそれぞれにつき，(1) 先行研究に基づく画像説明文生成モデルと (2) 中間表現と脳活動データとの対応関係の学習を行い，脳活動データ説明文生成モデルを構築した．

4.1　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] を利用した．

4.1.1　先行研究に基づく画像説明文生成モデル

train 用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．学習に関するハイパーパラメータは，学習率を 1.0（14,000 毎に ×0.999）とした他，Chainer で採用されている深層学習の効率化手法を取り入れ，勾配閾値 5，L2 正則化項 0.005 とした．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 512 次元の語彙とし，LSTM ユニット数は Attention 適用時は各層 14×14=196，非適用時は 1,000 に設定した．学習するパラメータは Attention Mechanism と Decoder の重み係数とし，[-0.1,0.1] でランダムに初期化した．Encoder は事前学習した VGGNet 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用いる．

4.1.2　中間表現と脳活動データとの対応関係

train 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）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う．train 用データ数は 3,600 個であり，画像のサイズは VGGNet の入力次元に揃え 224×224 とした．学習する 3 層 MLP のハイパーパラメータについては，学

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

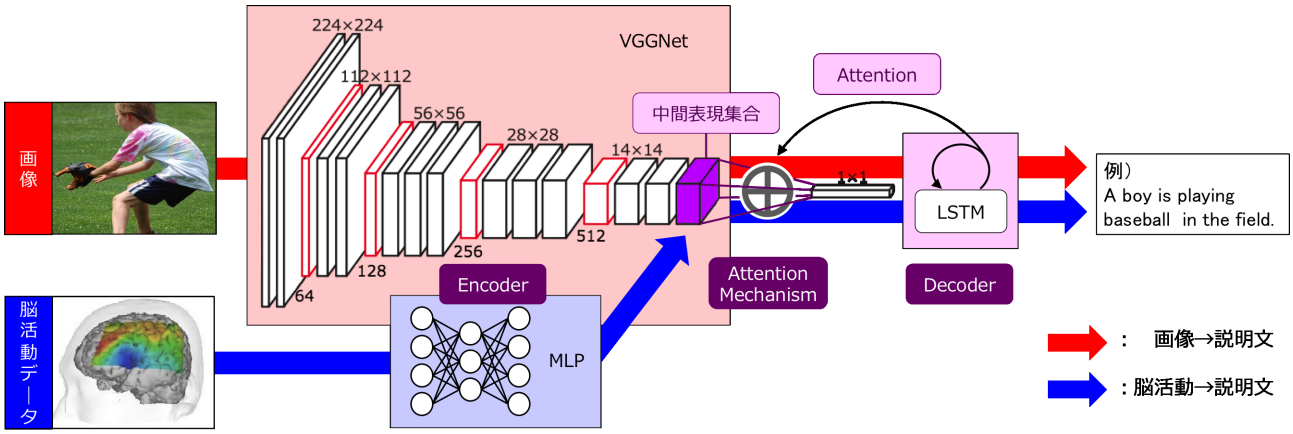

図 1: 本研究の概要図（Attention Mechanism を適用したモデル）

習率 0.01，勾配閾値 5，L2 正則化項 0.005，中間層ユニット数 1,000 に設定した．パラメータは [-0.2,0.2] 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．

4.2 実験 1：Attention Mechanism 適用モデル

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 2 に示す．また，表 1 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．

出力された説明文は文章として成立しておらず，画像の内容もあまり捉えられていない．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．

表 1: 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

A in of man people with and is on standing.

Man a people in of with and is on street.

図 2: 実験 1：生成した説明文とその時見ていた画像例

4.3 実験 2：Attention Mechanism 非適用モデル

実験 1 と同様，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 3 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 2 に示す．

Attention 適用時に比べ，文法的にも内容的にもより適切な表現を獲得しており，平均二乗誤差の減少にも見られるように，対応関係を学習する中間表現の次元が 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．

表 2: 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

A group of people sitting next to each other.

Top of a man room on front to the street.

図 3: 実験 2：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，MLP を用いて脳活動データと 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 Enc-DecNet による画像説明文生成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．

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 などの指標を用いた実験結果の評価および考察などが挙げられる．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

[1] O.Vinyals, A.Toshev, S.Bengio, D.Erhan, ”Show and tell: a neural image caption generator,” in CVPR’ 2015, 2015.
[2] 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.
[3] K. Cho, A. Courville, Y. Bengio. ”Describing Multimedia Content using Attention-based Encoder-Decoder Networks, ”CoRR, abs/1507.01053, 2015.
[4] K. Simonyan, A. Zisserman,”Very deep convolutional networks for large-scale image recognition,”in ICLR, 2015.