深層学習を用いた画像刺激下の脳活動データの分析と言語文生成

理学専攻・情報科学コース　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，刺激を受けた際の脳活動情報から人の想起している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想起し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し，言語表現を介した脳活動情報の分析を目指す．その際，脳活動データは取得のためのコストが大きく大量収集が困難なため，画像からそこに映る事象の説明文を生成するよう事前学習されたシステムを援用した．

2　脳活動データの説明文生成

本提案手法は，2 種類のニューラルネットワーク（NN）モデルを組み合わせ，人間の脳神経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行う．図 1 に概要図を示す．

2.1　(A) 画像→説明文モデル

Vinyals ら [1] は，Encoder，Decoder の役割を果たす 2 つの深層学習モデルを連結することで入力を中間表現となる数値ベクトルに encode し，再び decode して別の形に出力する Encoder-Decoder Network[2] を用いて，画像に映る事象を言葉で説明する画像説明文生成を行っている．本手法では，画像特徴量を中間表現とし，画像特徴抽出を行う CNN の一種である VGGNet[3] を Encoder，文生成を行う言語モデルとなる 2 層 LSTM-LM[2] を Decoder とした画像→画像特徴量→説明文モデルを構築する．なお，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．

2.2　(B) 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であり，本稿では 3 層 NN を用いた実験結果を示す，なお，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．

2.3　提案手法：(C) 脳活動データ→説明文モデル

上記の画像→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．具体的な処理手順としては，まず脳活動データ→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，以降は画像→説明文モデルの LSTM-LM で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単語出力を繰り返し，1 語ずつ出力して文章を生成する．

3　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．

表 1: 実験設定（詳細）

	(A) 画像→画像特徴量→説明文モデル	(B) 脳活動データ→画像特徴量モデル
データセット	Microsoft COCO	動画刺激による脳活動データ
学習量	414,113 sample×100 epoch	4,500 sample×1,000 epoch
アルゴリズム	Adam	stochastic gradient descent
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	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
層ユニット数	各層 512	65,665 - 8,000 - 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

3.1　実験 (A)：画像→画像特徴量→説明文モデル

学習用のデータとして 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．数値設定については画像説明文生成の先行研究に基づいて調整した．詳細は表 1 の左列に示す．

学習結果として epoch 毎に評価用画像からの出力文の perplexity を記録したところ，単調に減少し約 2.5 に収束したことより，学習の進度が確認できた．また，評価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 2 に示す．

考察として，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おり，文法も大きな崩れがなく細部の前置詞（in,on）や冠詞（a,an）も正確であることから，出力された説明文は十分に画像の大意を認識し表現できていると言える．学習に使われていない評価用の画像に対しても相応な説明文を生成でき，perplexity も収束していることから画像→説明文モデルについて適切な学習が行われたと評価できる．

A man is surfing in the ocean on his surfboard.

A black and white cat is sitting on the toilet.

図 2: 実験 (A)：画像から生成した説明文の例，画像はランダムに選出

3.2　実験 (B)：脳活動データ→画像特徴量モデル

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，functional Magnetic Resonance Imaging（fMRI）を用いて記録した被験者に動画像を見せた時の脳神経活動データ，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静止画を使用する．学習データ数は 4,500 で，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

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

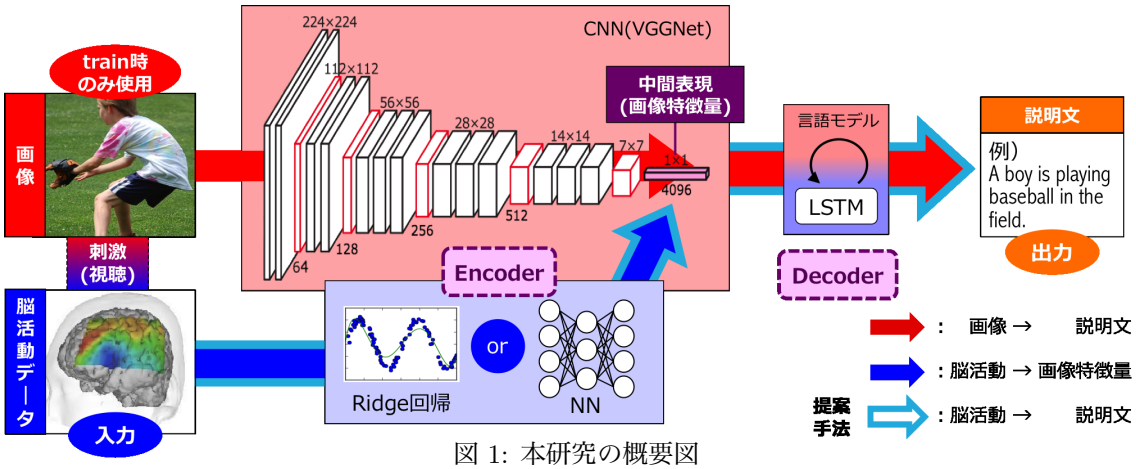

図 1: 本研究の概要図

を入力とし，その時見ている静止画から VGGNet により得られる 4,096 次元の画像特徴量との対応を 3 層 NN で学習する．その他詳細は表 1 の右列に示す．

学習結果として epoch 毎に平均二乗誤差を記録したところ，ほぼ単調に減少し約 1.13 に収束したことより，学習の進度を確認できた．また，評価のため，Microsoft COCO の学習データセットに含まれる 82,783 画像を VGGNet に入力して計算した画像特徴量のうち，このモデルが脳活動データから予測した画像特徴量に近いものを二乗誤差によって検索することで，脳活動データから刺激画像に類似した画像を取り出す実験を行った．図 3 に類似画像検索の結果を示す．

学習したモデルはほとんどの学習用画像刺激データに対して適切な画像を検索しており，脳活動データを入力として画像のデータベースから被験者への刺激画像に近い画像を示すこと，すなわち脳活動情報から画像特徴量を抽出できていることを確認できる．

図 3: 実験 (B)：刺激画像とそのときの脳活動データから計算された類似画像 (Top-3)

3.3 実験 (C)：脳活動データ→説明文モデル

実験 (A)，(B) で学習したモデルを組み合わせ，脳活動データからの説明文生成を実現する．

図 4 に示すように，学習用データから選んだ 2 つの脳活動データに対し，人間が解釈しうる説明文章が生成された．また，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ところ，脳活動データからの生成文と画像からの生成文の一致が多く見られたことより，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．

	刺激画像	脳活動→説明文モデル	画像→説明文モデル
train data		A man is surfing in the ocean on his surf board.	A man is surfing in the ocean on his surf board.
		A close up of an orange and white clock.	A pair of scissors sitting on the ground.
test data		A group of people standing next to each other.	A group of people standing next to each other.
		A man walking down the street with an umbrella.	A train traveling down tracks next to trees.

図 4: 実験 (C)：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，画像から生成した説明文の例

4 おわりに

本研究では，深層学習を用いた画像説明文生成システムを援用し，脳活動データと CNN による画像特徴量との対応関係を学習した 3 層 NN 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，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，ネットワーク構造の変更による精度向上が挙げられる．特に，脳活動情報の持つ時系列性に着目し，画像説明文生成手法に代わっての動画説明文生成手法の適用には検討の価値があると考えている．また，fMRI 脳活動データの単純な増加や複数被験者のデータを同時に使用する工夫を加えるなど学習データの増加が望まれる．更なる追加実験や機械的指標を用いた評価による脳活動データの更なる分析も行うべき課題である．

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

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

[4] S. Hochreiter, J. Schmidhuber, ”Long short-term memory,” Neural Computation, vol. 9, no. 8, pp. 1735-1780, 1997.