視覚刺激からの脳活動解読と状態推定

理学専攻・情報科学コース　張 嘉瑩　(指導教員：小林 一郎)

1　はじめに

近年，脳神経科学分野においてヒト脳内における意味表象の分析に関する研究が盛んに行われている．特に，自然言語処理分野における深層学習手法を取り入れた研究が増えている．一方で，大規模な脳活動データの収集は難しく，一度に大量のデータを必要とする深層学習手法を用いる際には困難が生じる．本研究では，異なる被験者間の脳活動データの形式を統一することで脳活動データのデータ拡張を目指す．また，動画刺激と画像刺激といった 2 つの視覚刺激からの脳活動データを対象に，人が視覚刺激によって頭の中に抱いた意味表象を自然言語文によって説明する深層学習手法を構築し，脳活動解読の観点からの分析も行う．

2　脳活動データの拡張

異なる被験者同士の脳活動データを統一する際に，マルチモーダル情報を相互に変換する手法を適用して，ある被験者の脳活動データを他の被験者の脳活動データから擬似的に作り出すことを考える．

2.1　概要

図 1 に提案手法の概要を示す．具体的には，使用する脳活動データを，動画像タスク 1 のみを視聴した被験者 A と動画像タスク 1 と 2 を視聴した被験者 B のものと仮定して，以下の手続きを行う．

step 1. 学習
Vukotic ら [3] によって提案された Bidirectional Deep Neural Network (BiDNN) を用いて，動画像タスク 1 を視聴したときの A と B の脳活動データをそれぞれ入出力とし，対応関係を学習する．

step 2. 変換
step 1. におけるモデルを使用し，学習に用いたタスク 1 とは別の動画像タスク 2 を視聴した B の脳活動データを，A の脳活動データに変換する．これにより，擬似的に動画像タスク 2 を視聴した A の脳活動データが得られる．

step 3. 評価
評価においては，松尾ら [1] が提案した手法を利用する．動画像タスクを視聴しているときの脳活動データを入力とし，その際，被験者が想起している意味情報の説明文を出力する．動画像タスク 1 のみを視聴した被験者の脳活動データを入力としたときと比べて，動画像タスク 1 と 2 の両方を視聴した被験者の脳活動データを入力としたときに出力された説明文の精度が上がっているのか確認する．

2.2　実験

使用するデータは，動画像タスクを被験者に視聴させたときの血中酸素飽和度信号 (BOLD 信号) を fMRI を用いて記録した脳活動データである．被験者 A は脳活動の観測領域のうち皮質に相当する 62,552 次元のデータを，被験者 B は 70,933 次元のデータを使用している．

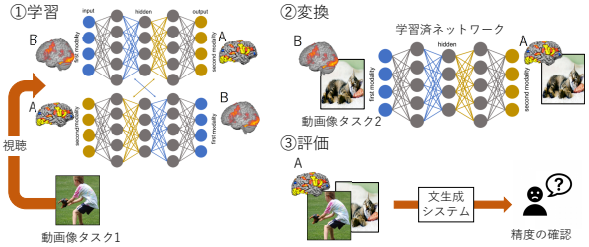

図 1: 脳活動データの拡張

評価において，まず，動画像視聴時に見ていたとされる動画像から切り出した静止画に対して caption 生成システムを用いて生成した説明文を正解文とする．次に，動画像視聴時の脳活動データを入力とし，その時想起している言語意味情報の説明文と先ほどの正解文を BLEU スコアで評価する．データ数は動画タスク 1 が 7,200 サンプル，動画タスク 2 が 9,000 サンプルである．評価実験における BLEU スコアを表 1 に示す．

表 1: 評価実験における BLEU スコア

被験者	タスク	BLEU スコア (train / test)
A	1	0.5026 / 0.5027
	1, 2(擬似)	0.5238 / 0.5099
B	1	0.5028 / 0.5085
	1, 2	0.5191 / 0.5086

被験者 A における評価実験の結果を見ると，擬似的なデータを増やした際にはわずかに BLEU スコアが上がっている．当初予想していたほどスコアは上がらなかったが，擬似的ではなく実際にタスク 2 を視聴した被験者 B における結果でも BLEU スコアが大きくは上がっていない．その原因として，使用したデータが視聴している動画に直感的に説明文をつけることが難しいものが含まれているためだと考えられる．

3　動画刺激からの脳活動データ

松尾ら [1] 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．

3.1　概要

図 2 に提案手法の概要を示す．具体的には，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
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
Venugopalan ら [2] によって提案されたモデル (S2VT) を使用する．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このモデルを学習する．

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．

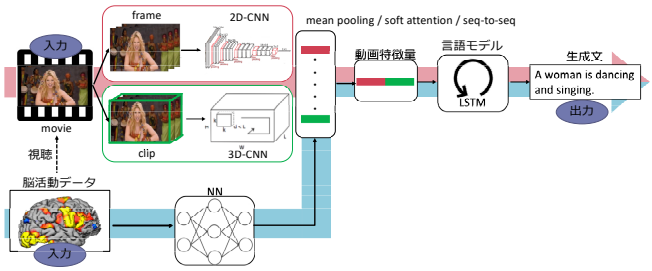

図 2: 動画刺激時の脳活動データからの文生成

3.2 実験

step 1. の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) を使用する．学習に関する詳細設定は表 3 の左列に示す．結果は，表 2 に示す．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捕らえられた可読性のある自然言語文の生成ができたと言える．

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.

step 2. と step 3. の脳活動データは皮質に相当する 62,552 次元のみを使用する．学習に関する詳細設定は表 3 の右列に示す．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 S2VT モデルを使用した説明文生成のいずれもほとんどの出力文が同じもの (例 : A person is cleaning the floor.) になった．原因としては，S2VT モデルを学習した際の動画データセットと被験者が視聴した動画の種類が大きく離れているためだと考えられる．

表 3: 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

4 画像刺激からの脳活動データ

先行研究 [1] に従ってモデルを構築し文生成をする．使用するデータセットと一部モデルを変更することにより，更なる分析を行う．

4.1 BOLD5000

被験者に画像を視聴させた時の血中酸素飽和度信号を fMRI を用いて記録した BOLD5000 データセットを使用する．画像の種類は，Scene，Microsoft COCO，ImageNet といった画像を用いた研究で広く使われているものである．データ数は 4,916 であり，皮質に相当する 43,312 次元を使用する．

4.2 実験

先行研究 [1] に従ってモデルを構築し，動画像刺激時のものから画像刺激時の BOLD5000 へのデータセット変更と誤差関数の変更をして実験を行う．画像刺激時のデータを使用することにより，脳活動データとその時見ていたとされる画像の対応づけがより明確になると言える．文生成の例を表 4 に示す．train については概ね画像の特徴を捉えた文章が出力されたが，test については文法は正しいが特徴はまだ捉えられていない．

表 4: 脳活動から生成した文の例 (上 : train，下 : test)

A brown bear standing in the grass near some trees.

A dog sitting on the ground next to an umbrella.

5 おわりに

本研究では，脳活動データのデータ拡張を行うと共に，視覚刺激時の脳活動データを対象に脳活動解読の観点から分析を行った．データ拡張については，拡張したデータと実データによる結果をそれぞれ比較し，結果が似ていることを確認した．動画刺激時の脳活動データについては，脳活動データからその時被験者が想起していたとされる意味表象を文として出力するモデルを構築して分析を行った．画像刺激時の脳活動データについては，誤差関数を変更することにより，train データについてより良く特徴を捉えられた文を出力できることを示した．

参考文献

[1] 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.

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

[3] V. Vukotic, C. Raymond and G. Gravier. Bidirectional Joint Representation Learning with Symmetrical Deep Neural Networks for Multimodal and Crossmodal Applications. ACM International Conference on Multimedia Retrieval (ICMR), 343-346, 2016.