深層学習を用いた系列データからのテキスト生成による事象の理解

理学専攻・情報科学コース　漆原 理乃　(指導教員：小林 一郎)

1　はじめに

近年, 深層学習を用いたテキスト生成の研究が盛んに行われている. 本研究では, 系列データ内の事象を表現するテキストの生成を目的とし，対象となる事象が異なる 2 つのモデルを構築した. 1 つ目は動画像中の事象を捉えた動画像説明文生成モデルである. 2 つ目は音声刺激下のヒトの脳活動情報の言語解読モデルである. 音声刺激下の脳活動情報から, ヒトが脳内に想起した高次意味表象を言語として生成する手法を構築した.

2　動画像説明文生成モデル

2.1　モデル概要

図 1 に本研究で提案する動画像説明文生成モデルの概要図を示す. 動画像のフレームごとに人の姿勢情報を抽出し時系列情報として，動作を表す単語を選択する処理と，フレームごとに物体を検出する処理を合わせ，それぞれの処理において得られた結果から人の動作を捉えた動画像説明文生成を行う.

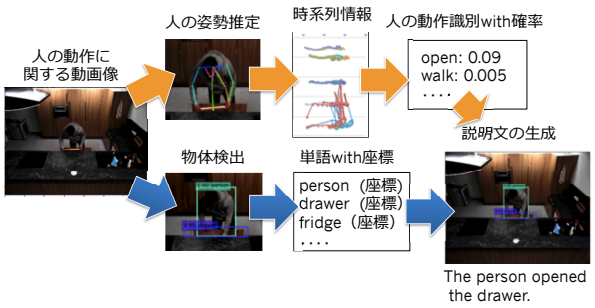

図 1: 動画像説明文生成モデルの概要図

2.1.1　動作認識

本研究では，Cao らによる深層学習を用いた人の姿勢推定手法 [1] を用い, 動画像の各フレームごとに鼻や目, 肘などの 18 個の人の部位の座標を検出する. そこで得られたフレームごとの 36 次元の情報 (=人の部位数 18× フレームの座標数 2) から, Encoder-Decoder Temporal Convolutional Networks(ED-TCN) [2] を用いて, 広範囲の時系列情報を効果的に捉え, 動画像中の全てのフレームに対して動作を表す適切な単語を選択する.

2.1.2　物体検出

物体検出には，Single Shot MultiBox Detector(SSD) [3] を用いる. SSD は, 画像を入力とし, 画像中に含まれる物体の種類とその物体の座標 {x 軸の最大値, x 軸の最小値, y 軸の最大値, y 軸の最小値 }, 確信度を出力する.

2.1.3　文生成

Sutskever らによる Long Short-Term Memory(LSTM) を用いた言語モデル (sequence-to-sequence) [4] を改良し, 動画像中の物体の位置情報を用いた文生成手法を提案する. 動画像の各フレームごとに ED-TCN によって予測された動作を表す単語と, SSD によって検出された物体の単語と検出された物体それぞれの位置情報となる座標 {x 軸の最大値, x 軸の最小値, y 軸の最大値, y 軸の最小値 } を入力とし, すでに学習されたモデルから物体の位置情報や語順情報に基づき, 各語が選ばれる確率を算出し, 逐次的に次の単語の予測することで文を生成する.

2.2　実験

2.2.1　実験設定

料理の動画像とフレームごとに対応する説明文からなるデータセットである TACoS Cooking Dataset を使用した. 説明文は提案手法に合わせ, 時制を過去形に一致させ, 位置情報を追加し, 単語を抽象化するなど, 手動で改変したものを使用した. 動作識別においては, 動画像の訓練データで使用された 58 個の動作を表すフレーズを使用した. 物体検出においては, 大規模な画像認識コンペティション VOC2007 における 20 種類と TACoS の説明文において頻繁に使用される 13 種類の単語の合計 33 種類の物体を検出した.

2.2.2　実験結果

評価用の動画像において SSD の結果と生成した文を図 2 に示す. 図 2 の生成文から, (1) では人は fridge の近くにおり, (3) では cupboard の近くにいるため, それらの単語が文中に出現したと考えられる. また, 人の動作を表す took out や moved も文中に出現し, 人の動作も適切に予測できることも確認できた.

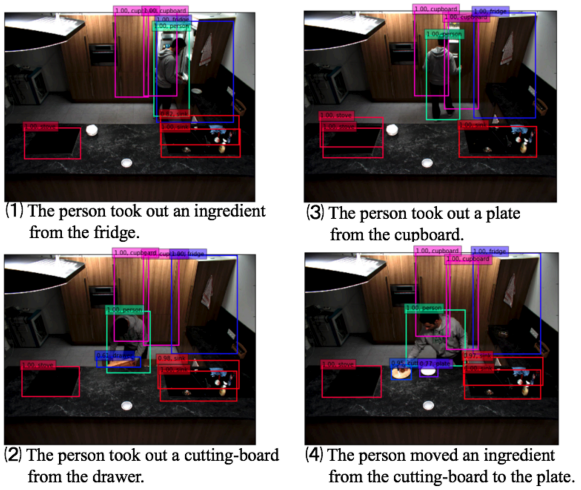

図 2: 物体検出結果と生成文

3　脳活動情報の言語解読モデル

3.1　モデル概要

深層学習を用いて, 音声刺激を受けた脳活動データを入力とし, その時に刺激となっていた音声のテキストを生成することで, ヒトが頭の中で想起した意味表象を

言葉として解読することを目指す．しかし，Functional Magnetic Resonance Imaging(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，Encoder-Decoder Network に基づく自動音声認識手法を援用することで少量データを効率的に活用する．図 3 に本提案手法の概要図を示す．また，実行時の手順を示す．

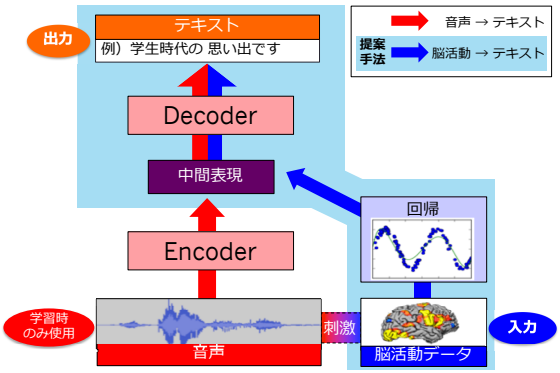

図 3: 脳活動情報からのテキスト生成モデルの概要図

step 1. 自動音声認識
Encoder-Decoder Network を用いた自動音声認識モデル，Hybrid CTC/Attention Architecture [5] を使用．

step 1-1. Encoder: 音声中間表現の抽出
自動音声認識の Encoder を用いて，音声から音声中間表現を抽出．

step 1-2. Decoder: テキスト生成
step1-1. において抽出された音声中間表現を，自動音声認識の Decoder に入力し，テキストを生成．

step 2. 脳活動データの特徴量推定
脳活動データとその刺激である音声の中間表現 (step1-1 の出力) との対応関係を学習した回帰モデルにより，新規の脳活動データから対応する音声中間表現を推定．

step 3. 脳活動データの特徴量からテキスト生成
step1-2. で学習済みの自動音声認識の Decoder を用いて，step2. で計算された新規の脳活動データの特徴量を入力として，テキストを生成．

3.2 実験

3.2.1 実験設定

step1. の自動音声認識では，深層学習を用いた音声認識ツールキッド ESPnet を使用し，データセットとして日本語話し言葉コーパス (CSJ) の 1 本 10 分程度の講演データ 3,254 本を使用した．step2. の脳活動データの特徴量推定では，Ridge 回帰を用い，データセットとして CSJ の 16 本を 1 人の被験者に聴かせた時の血中酸素濃度依存性信号 (BOLD 信号;Blood Oxygenation Level Dependent Signal) を fMRI を用いて 1 秒ごとに記録した脳活動データ，および fMRI のデータ収集と同期させた CSJ の音声を使用した．立体撮像 96 × 96 × 72 ボクセルのうち大脳皮質と，大脳皮質中の音声処理を行う脳領域である前頭葉，頭頂葉，側頭葉に相当するデータ列を用い，fMRI の観測遅延を考慮し 4,5,6 秒後の脳活動データを Ridge 回帰の説明変数とした．

表 1: 脳領域ごとの step.2 の結果

脳領域 (データの種類)	相関係数
大脳皮質 (訓練データ)	0.99
大脳皮質 (評価データ)	0.41
前頭葉 (評価データ)	0.26
頭頂葉 (評価データ)	0.22
側頭葉 (評価データ)	0.25

表 2: 大脳皮質を用いた step3. の結果

データセットの種類	正解テキスト	生成テキスト
訓練データ	だんだんと興味が	だんだんと興味が
訓練データ	できました	できました
評価データ	あの楽しい	映画
評価データ	ですがも	です

3.2.2 実験結果

step2 において，ridge 回帰の学習時に使用した訓練データ，使用していない評価データ，それぞれの脳活動データの特徴量を推定し，推定結果とその刺激である音声の中間表現との相関係数を計算した結果を表 1 に示す．表 1 から，評価データを比較すると特定の脳領域より大脳皮質全体を使用する方が相関係数が高いことがわかる．訓練および評価データの比較においては，後者においては高い相関係数を得ることができておらず，これはデータ数が少ないことに起因していると考えられる．また，step3 におけるテキスト生成結果を表 2 に示す．訓練データからは正解と同じテキストを生成することができた．評価データで生成できていないのは，step2 のモデルの精度が低いためと考えられる．

4 おわりに

本研究では，2 つのテキスト生成モデルの構築を行った．動画像説明文生成モデルでは，時系列データにおける人の動作や位置情報を踏まえたテキスト生成ができていることを確認した．また，脳活動情報の言語解読モデルにおいては，訓練データにおいては言語解読が可能であることを確認し，特定の領域を入力とした比較実験から，人間の脳における音声認識処理が特定の領域より脳全体で行われていると示唆することができる．

参考文献

[1] Z. Cao, T. Simon, S. Wei, and Y. Sheikh. “Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields”, In CVPR, 2017.

[2] C. Lea, M. D. Flynn, R. Vidal, A. Reiter, and G. D. Hager. “Temporal Convolutional Networks for Action Segmentation and Detection”, arXiv preprint arXiv:1611.05267, 2016.

[3] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. Fu, and A. C. Berg. “SSD: Single Shot MultiBox Detector”, arXiv preprint arXiv:1512.02325, 2016.

[4] I.Sutskever, O. Vinyals, and Q. V. Le. “Sequence to sequence learning with neural networks”, In Advances in NIPS, 2014.

[5] S. Watanabe, T. Hori, S. Kim, J. Hershey, T. Hayashi. “Hybrid CTC/attention architecture for end-to-end speech recognition.” IEEE Journal of Selected Topics in Signal Processing 11.8 (2017): 1240-1253.