人の動作および物体認識に基づく動画像からの文生成

漆原 理乃　(指導教員：小林 一郎)

1　はじめに

近年、監視カメラによる不審者の挙動の把握や高齢者の見守りなど，人の動作を言葉によって報告する技術の必要性が高まっており，深層学習を用いた動画像や画像の言語化手法は多く研究されているが，人の動作を正しく捉えて言語化する手法はほとんどない．そのため本研究では，深層学習を用いて動画像中の事象を正しく捉えた説明文生成に取り組む．図 1 に本研究の概要図を示す．具体的には，動画像のフレームごとに人の姿勢情報を抽出し時系列データとして，動作を識別する処理と，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や物体を正しく捉えた説明文生成を行う．

2　人の動作認識に基づく文生成

2.1　動作識別

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 個の人の部位のピクセル座標を検出する．そこで得られたフレームごとの 36 次元 (= 人の部位数 18× フレームのピクセル座標数 2) の情報に対して，Encoder-Decoder Temporal Convolutional Networks (ED-TCN) [2] を用いて，プーリングとアップサンプリングにより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．

2.2　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，VGG16 をネットワーク構造のベースとし，画像中の物体を検出するシステムである．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ピクセル座標 {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．

2.3　物体の位置情報を用いた文生成

本研究では，機械翻訳手法の 1 つである，Sutskever ら [4] による Long Short-Term Memory (LSTM) を用いた，あるシーケンスを別のシーケンスに変換する言語モデルを改良し，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．図 2 に提案手法のモデルを示す．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 (図 2 では verb) と，SSD によって検出された物体の単語 (図 2 では w1，w2) と検出された物体それぞれの位置情報となるピクセル座標 {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ピクセル座標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を繰り返し文を生成する手法を提案する．

3　実験

3.1　使用データ

料理の動画像である TACoS Cooking Dataset [5] とそのフレームごとに対応する説明文である TACoS Multi-Level Corpus [6] を使用した．

3.2　実験設定

3.2.1　動作識別における実験設定

人の姿勢推定の実装に際しては，[1] のコード[1]を深層学習フレームワーク TensorFlow を用いて実装した．ハイパーパラメータの数値設定やネットワークの重みは Microsoft COCO で学習済みのものを使用した．

ED-TCN に関しては，深層学習フレームワーク Keras で実装されているコード[2]を TensorFlow のバックグラウンドのもと使用した．実験に使用した動画像は，10fps で 40 秒から 22 分程度 (フレーム数は 363 から 13,648) の 121 本で，訓練用に 109 本，検証用に 6 本，評価用に 6 本使用した．ED-TCN の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 [2] における実験同様，レイヤー数は Encoder と Decoder それぞれ 2 層使用し，フィルタサイズは Encoder 側では第 1 層目は 64 と第 2 層目は 96 に設定し，Decoder 側の最終層は 64，最終層の 1 つ前の層は 96 とする．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，全てのレイヤーの畳み込み層においてドロップアウトを使用し，500 epochs で，時系列に畳み込むフレームサイズは 20 として実験を行った．

3.2.2　物体検出における実験設定

物体検出手法 SSD では深層学習のフレームワーク Keras によって実装されているコード[3]を用いた．SSD で使用されている VGG16 のネットワーク構造のうち，画像特徴量を抽出する層である，conv1_1 から pool3 までの層においては，大規模な画像認識コンペティション VOC2007 の 5,011 枚の画像のみで学習した重みを使用した．その層以降 (つまり conv4_1 以降) の学習は TACoS の全動画像からランダムに抽出した 200 枚のフレーム画像も交え 5,211 枚の画像で学習した．検出する物体の種類は VOC2007 における 20 種類と，TACoS において説明文で頻繁に使用される代表的な 13 種類の単語を厳選し，合計 33 種類の物体を画像から検出するように学習した．

3.2.3　文生成における実験設定

文生成システムの実装に関しては，言語モデル [4] を深層学習フレームワーク Keras で実装したコード[4]を用いた．入力は，SSD で検出できる 33 種類の単語と訓練用の動画像の説明文で使用されている単語を合わせた 63 単語のうち，SSD で検出した物体の単語は 1

[1] https://github.com/ZheC/Realtime_Multi-Person_Pose_Estimation
[2] https://github.com/colincsl/TemporalConvolutionalNetworks
[3] https://github.com/rykov8/ssd_keras
[4] https://github.com/farizrahman4u/seq2seq

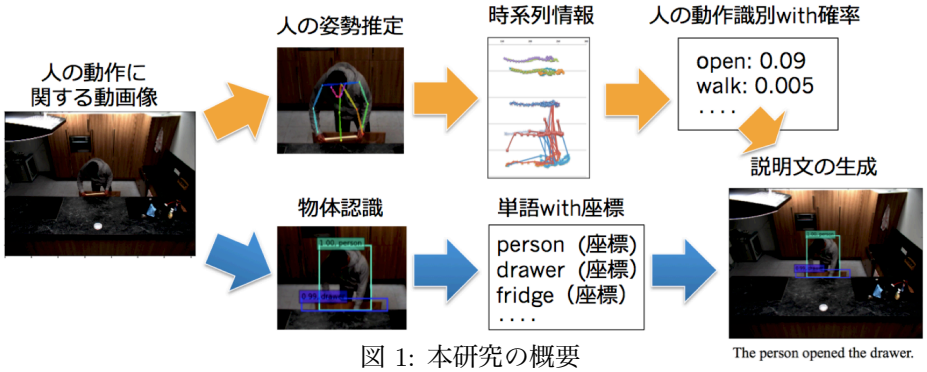

図 1: 本研究の概要

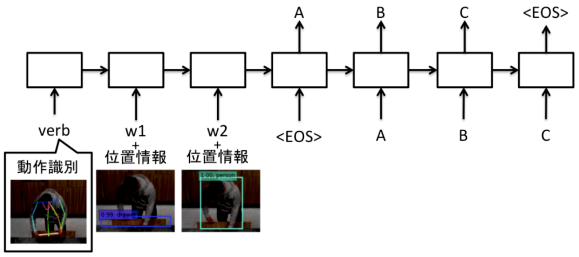

図 2: 物体の位置情報を用いた文生成モデル

でそれ以外を 0 としたベクトル表現に，検出した物体の位置情報 {x 軸の最大値，x 軸の最小値，y 軸の最大値，y 軸の最小値 } を追加する．つまり，入力は 67 次元で出力は 63 次元とした．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，100 epochs で実験した．表 1 に実験で使用した動画像の概要を示す．複数のフレームで 1 つの文が当てられており，同一の文でもフレームが異なれば，SSD で検出した物体の位置情報が異なるため，5 つの訓練用の動画像における 112 文と 5,715 枚の画像で学習を行った．

表 1: 使用した動画像の概要

動画像 ID	種類	文数	フレーム数
s13-d21	訓練	13	702
s13-d25	訓練	13	716
s13-d28	訓練	30	1389
s13-d40	評価	16	901
s13-d52	訓練	16	713
s13-d54	訓練	40	2195

3.3 実験結果

TACoS のあるフレーム画像に対して姿勢推定を行った結果を図 3 に示す．評価用の動画像において SSD の結果と生成した文を図 4 に示す．また文生成における定量的な結果として，epochs ごとの評価用動画像の全フレームの BLEU スコアのマクロ平均を表 2 に示す．

表 2: 文生成における BLEU スコアのマクロ平均

epoch 数	BLEU スコア
50 epochs	0.70
100 epochs	0.71

3.4 考察

表 2 からはある程度高精度で文が生成されていることがわかる．図 4 の具体的な生成文を見てみると，(1) では人は fridge の近くにおり，(2) 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，SSD の結果では現れていない単語 ((1)ingredient，(2)plate など) も文中に出現し，訓練用データから適切に補完できていることも確認できる．

図 3: TACoS のあるフレームにおける姿勢推定結果

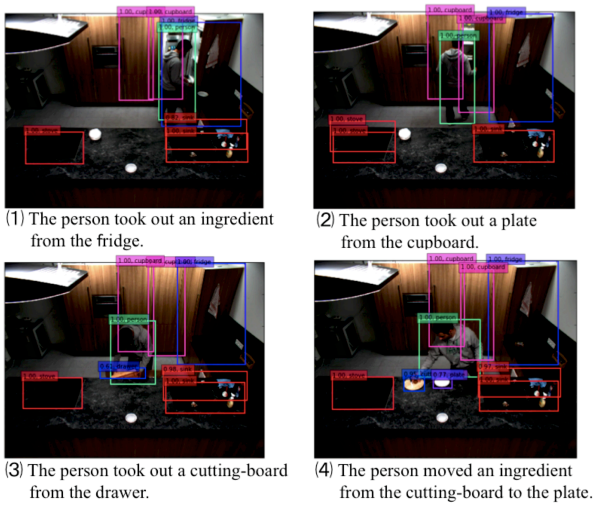

図 4: SSD 結果と生成文の例

4 おわりに

本研究では，動画像のフレームごとに人の姿勢情報を抽出し時系列データとして，ED-TCN を用いて動作を識別する処理と，SSD を用いたフレームごとに物体検出を行う処理を合わせ，動画像中の事象を正しく捉えた説明文生成手法を構築した．実験により，識別した動作と物体の位置情報や語順を踏まえて文が生成されていることを確認した．

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] M. Regneri, M. Rohrbach, D. Wetzel, S. Thater, B. Schiele, and M. Pinkal, “Grounding action descriptions in videos,” Transactions of the Association for Computational Linguistics (TACL), vol. 1, pp. 25-36, 2013.

[6] A. Senina, M. Rohrbach, W. Qiu, A. Friedrich, S. Amin, M. Andriluka, M. Pinkal, and B. Schiele. “ Coherent multi-sentence video description with variable level of detail”, arXiv preprint arXiv:1403.6173, 2014.