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 †　　小林 一郎 ‡

† お茶の水女子大学 理学部 情報科学科　　‡ お茶の水女子大学基幹研究院

{g1420509, koba}@is.ocha.ac.jp

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，スポーツの実況中継など，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，深層学習を用いた画像の言語化に関する研究が盛んに行われている [1, 2]．深層学習を用いて画像中の物体を検出し，得られた単語から確率的に説明文を生成する手法 [3] も報告されている．また，動画像における言語化手法としても、Encoder-Decoder Network に基づく研究 [4] が盛んで，特に動画像処理に関する様々なタスクにおける動画像特徴量抽出に効果的な Convolutional 3D [5] を用いて説明文生成を行う手法 [6] なども報告されている．ただ、これらは深層学習を用いるため新規な自然文を生成することに成功しているが，画像や動画像特徴量から Decoder を介して文章を生成する手法が多く，人間が実際に画像や動画像を見て認識するような事象，特に人の動作について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動画像の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．

2 人の動作認識に基づく文章生成

2.1 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 [7] 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) [8] を用いて，動画像のフレームごとのセンサー情報や特徴量を入力とし，プーリングとアップサンプリング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．ED-TCN は機械学習分野における代表的なラベリング手法であるサポートベクトルマシン (SVM) や時系列データ対応の深層学習モデル Recurrent Neural Network (RNN) 用いたラベリング手法よりも，行動認識や行動セグメンテーションにおいて高精度かつ高速度である [8]．

2.2 物体検出

物体検出には Single Shot MultiBox Detector (SSD) [9] を用いる．SSD は，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．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 のネットワーク構造をベースとし，深層学習を用いた他の手法 Faster R-CNN や YOLO よりも高精度かつ高速度である [9]．

2.3 物体の位置情報を用いた文生成

本研究では，Sutskever ら [10] による RNN の一種である Long Short-Term Memory (LSTM) を用いた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 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

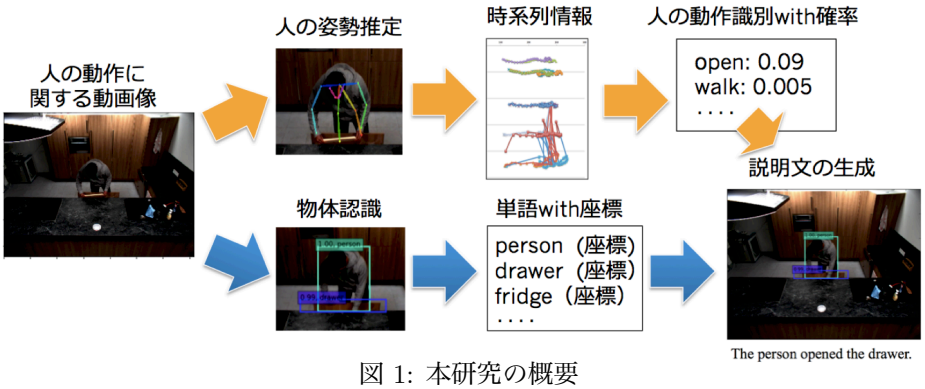


図 1: 本研究の概要

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．

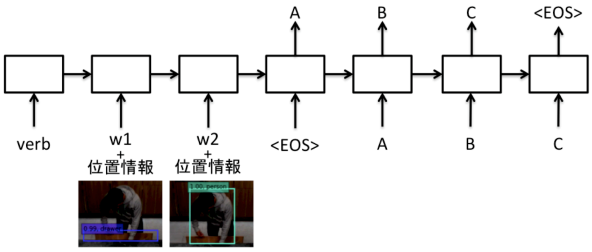


図 2: 物体の位置情報を用いた文生成モデル

3 実験

3.1 使用データ

料理の動画像である TACoS Cooking Dataset [11] とそのフレームごとに対応する説明文である TACoS Multi-Level Corpus [12][1]を使用した．フレームレートは 10 fps とし，40 秒 (フレーム数は 363) から 4 分 (フレーム数は 2,296) 程度の長さの 68 本の動画像とその説明文を用いた．

3.2 人の姿勢推定に基づく動作認識実験

本研究では，まず動画像の各フレームにおいて姿勢推定手法 [7] により人の姿勢を推定し，得られた時系列データを入力として，TCN を用いて人の動作を表す適切な単語を選択する．

3.2.1 実験設定

システムの実装に際しては，姿勢推定手法 [7] のコード[2]を深層学習フレームワークである TensorFlow を用いて実験を行った．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[3]で学習済みのものを使用した．

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[4]を TensorFlow のバックグラウンドのもと使用した．本研究では TACoS の動画像 68 本のうち 54 本を訓練用，6 本を検証用，8 本を評価用に使用した結果を提示する．また，訓練用データで使用された 46 個の動作を表す単語を識別に使用する．ここで，TACoS の動画像説明文中に動作を表す単語がないフレームもいくつかあり，その場合は None として実験した．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 [8] 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．時系列方向においてはフレームサイズを d とした場合に時刻 $t - \frac{d}{2}$ から時刻 $t + \frac{d}{2}$ までのフレーム情報を畳み込む手法 ([8] においては acausal と呼ばれる) を用い，200 epochs とし，畳み込むフレームサイズ d を変化させ実験した．

3.2.2 実験結果

TACoS のあるフレーム画像に対して [7] を行った結果を図 3 に示す．TCN において畳み込むフレームサイズ d を変化させ，実験を行った結果を表 1 に示す．評価手法としては，各動画像のフレームごとの正解率 (正解したフレーム数/全体のフレーム数) のマクロ平均と，時系列情報のセグメンテーションを評価する手法として [8] において使用されていた，編集距離を 0 から 100 に正規化した編集スコア (数値が高い方が良

[1]https://github.com/qiuwei/videosum/tree/master/corpus/TACoS
[2]https://github.com/ZheC/Realtime_Multi-Person_Pose_Estimation
[3]http://mscoco.org/
[4]https://github.com/colincsl/TemporalConvolutionalNetworks

い結果となる) をそれぞれ求め評価を行った．また，時系列方向に畳み込むフレームサイズ d を 20 とした場合の，46 単語のうちの一部単語において，それぞれの適合率，再現率，F 値を求め比較した結果を表 2 に示す．表 2 中の動画像数は訓練用データと検証用データに含まれる動画像の中で，その単語が出現する動画像数である．

図 3: TACoS のあるフレームにおける姿勢推定結果

表 1: 時系列方向の畳み込むフレームサイズの比較

畳み込みサイズ d	正解率	編集スコア
10	31.0	42.0
20	50.0	51.8
30	40.7	43.7

表 2: 単語ごとの比較

単語	適合率	再現率	F 値	動画像数
take out	0.59	0.71	0.65	60
wash	0.31	0.98	0.47	50
cut	0.14	0.08	0.10	47
peel	0.44	0.39	0.41	21
tap	0.0	0.0	0.0	2

3.2.3 考察

実験結果より，評価用の動画像において平均 50 % のフレームが動作を表す適切な単語を選択できることが確認できた．表 1 より 20 フレーム分 (2 秒程度) 畳み込む場合が正解率，編集スコアともに結果がよいこともわかる．各単語で比較してみると，表 2 より，その単語が含まれている動画像数が多い方が F 値は高く，少ないと低くなる．take out や wash は再現率は高く適合率は低いため，異なる動作の場合でもその 2 つを多く出してしまっており，cut は peel や slice などの似ている動作が多いため F 値が低いと考えられる．動作を表す単語選択においては似ている動作を識別できるような枠組みが必要と考えられる．

3.3 物体検出実験

物体検出手法 SSD において動画像の各フレームにおいて物体の種類とその位置情報の検出を行った．

3.3.1 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[5]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 の動画像からランダムに抽出した 251 枚のフレーム画像も交え 5,262 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 11 種類の単語を厳選し，合計 31 種類の物体を画像から検出するように学習した．

3.3.2 実験結果

SSD を実験した結果は図 4 に，3.4.2 項で生成した文と合わせて明記する．画像中には，検出された物体をバウンディングボックスで囲み，その上部左に検出した物体の種類の確信度を 0 から 1 の範囲に数値化したものを，上部右に検出した物体の種類を明記した．

3.4 文生成実験

物体検出手法 SSD において得られた情報と，動作認識実験において得られた動作を表す単語を用いて，文生成を行った．

3.4.1 実験設定

文生成においては，SSD において検出された物体の確信度が 0.6 以上の物体の単語と位置情報を入力として使用した．使用した動画像は TACoS の 3 本で 2 本を訓練用，1 本を評価用に使用した．表 3 に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いる．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．そのため 2 本の動画像における 43 文と 2,105 枚の画像で学習を行った．

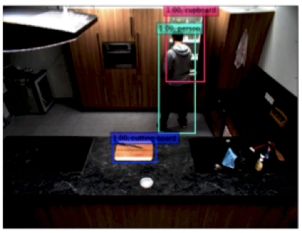
表 3: 使用した動画像の概要

動画像	種類	文章数	フレーム数
s13-d21	評価用	13	702
s13-d25	訓練用	13	716
s13-d28	訓練用	30	1,389

システムの実装に関しては，言語モデル [10] を深層学習フレームワーク Keras で実装したコード[6]を

[5]https://github.com/rykov8/ssd_keras
[6]https://github.com/farizrahman4u/seq2seq

(1) The person took out a cutting-board from the drawer.



(2) The person took out a plate from the cupboard.

(3) The person moved ingredients from the cutting-board to the plate.

(4) The person took out a tool from the drawer.

図 4: SSD 結果と生成文の例

用いた．入力は SSD で検出できる 31 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 49 次元のベクトル表現で，SSD において検出した物体の単語は 1 でそれ以外を 0 としたものと，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 } を入力次元に追加し，入力は 53 次元で出力は 49 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，20 epochs で実験した．

3.4.2 実験結果

評価用の動画像において SSD を実験した結果と生成した文を図 4 に示す．

3.4.3 考察

図 4 の生成文を見ると位置情報や語順などを踏まえて文が生成されていることが確認できる．特に図 4 (2) の生成文は，cutting board を認識しているが person からは離れた位置にあるため，文章中には cutting board は出現していない．それに対して，cupboard は person と近い位置にいるため，文中に出現したと考えられる．ただ，図 4 (1) の文章中に出現した単語 cutting board や，図 4 (2) の文章中に出現した単語 plate などは SSD の結果では現れていないため，訓練用データに特化していることも考えられる．

4 まとめと今後の課題

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，TCN を用いた動作を表す単語を選択する処理と，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，人の動作を捉えた説明文生成手法を構築した．人の姿勢推定に基づく動作認識実験においては，平均 50 %のフレームが動作を表す適切な単語を選択できることが確認できた．また，文生成実験においては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．

今後の課題としては，動作認識において cut や peel などの動作が近い単語でも識別できるような枠組みを考えていきたい．また，文生成実験において使用する動画像を増やし実験を行い，生成された文に対して BLEU などの手法を用いた定量的な評価も行っていきたい．

参考文献

[1] R. Kiros, R. Salakhutdinov, R. S. Zemel, “Unifying visual-semantic embeddings with multimodal neural language models”, In NIPS Deep Learning Workshop, 2015.

[2] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, “Show and tell: A neural image caption generator”, In CVPR, 2015.

[3] H. Fang, S. Gupta, F. Iandola, R. K. Srivastava, L. Deng, P. Dollar, J. Gao, X. He, M. Mitchell, J. C. Platt, C. L. Zitnick, and G. Zweig.“From captions to visual concepts and back”, In CVPR, 2015.

[4] R. Pasunuru, and M. Bansal, “Multi-Task Video Captioning with Video and Entailment Generation”, In ACL, 2017

[5] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, “Learning Spatiotemporal Features with 3D Convolutional Networks”, in ICCV, 2015

[6] J. Dong, X. Li, W. Lan, Y. Huo, C. G. M. Snoek, “Early Embedding and Late Reranking for Video Captioning”, In ACMMM, 2016

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

[8] 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.

[9] 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.

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

[11] 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 UTF201336, 2013.

[12] 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.