The 32nd Annual Conference of the Japanese Society for Artificial Intelligence, 2018

人の動作，物体検出およびそれらの位置情報を考慮した動画像からの文生成

Generating Descriptions for Videos based on the Interaction between a Human and Objects considering their Positional Relationship

漆原 理乃 *1　小林 一郎 *1
Rino Urushihara　Ichiro Kobayashi

*1お茶の水女子大学
Ochanomizu University

Image and video captioning with deep learning has been actively studied. On the other hand, few studies have focused on generating descriptions by correctly capturing the contents of images and videos. So, our study aims to generate natural language descriptions for videos based on the results of the identification of human activity and object detection for a movie. We propose a method to generate descriptions by combining four representative deep learning techniques: human pose estimation, time-series data analysis, object detection, and sequence-to-sequence. In addition, in generating descriptions, we apply positional information of the detected objects and a human in a movie, which works as natural attention on which objects should be described in a sentence. Through experiments, we have confirmed that our method could precisely describe the contents of movies, especially the interaction between a human and objects.

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いる．深層学習を用いた画像の言語化においては Encoder-Decoder Network に基づく研究 [Donahue 15, Kiros 15] が多く報告されており，Encoder として画像特徴量抽出に効果的な GoogleLeNet [Loffe 15] を用いた [Vinyals 15] が特に高性能な文生成手法を確立した．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 [Fang 15] 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 [Pasunuru 17] 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 [Tran 15] を用いて説明文生成を行う手法 [Dong 16] 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 [Cao 17] 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) [Lea 16] 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．

2.2 物体検出

物体検出には Single Shot MultiBox Detector (SSD) [Liu 16] 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 [Ren 16] や YOLO [Redmon 16] よりも高精度かつ高速度である [Liu 16].

2.3 物体の位置情報を用いた文生成

本研究では，Sutskever ら [Sutskever 14] による RNN の一種，Long Short-Term Memory (LSTM) を用いて，あるシーケンスを別のシーケンスに変換するという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 [Regneri 13] とそのフレームごとに対応する説明文である TACoS Multi-

連絡先: 漆原理乃，お茶の水女子大学理学部情報科学科小林研究室，〒 112-8610 東京都文京区大塚 2-1-1，g1420509@is.ocha.ac.jp

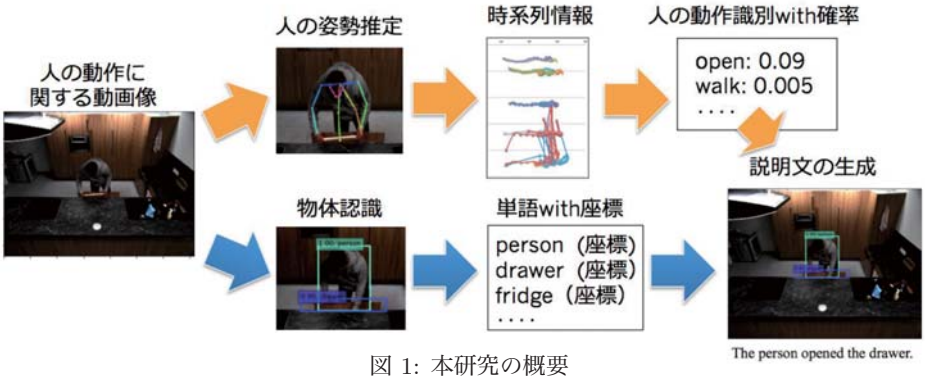

図 1: 本研究の概要

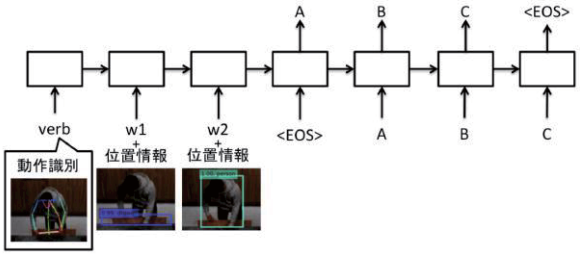

図 2: 物体の位置情報を用いた文生成モデル

Level Corpus [Senina 14][*1] 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，改変したものを使用した．

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 [Cao 17] 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，ED-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 [Cao 17] 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．

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 [*4] を TensorFlow のバックエ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 本使用した結果を提示する．また，訓練用データで使用された 58 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．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 [Lea 16] 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た．畳み込む手法や畳み込むフレームサイズ、epoch 数を変更させ実験を行った．ここで、畳み込む手法とは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 ([Lea 16] においては acausal と呼ばれる) と時刻 $t-d$ から時刻 t までのフレーム情報を畳み込む手法 ([Lea 16] においては causal と呼ばれる) の 2 種類があり，それぞれ実験をした．

3.2.2 実験結果

TACoS のあるフレーム画像に対して [Cao 17] を行った結果を図 3 に示す．ED-TCN において畳み込む手法とフレームサイズ d を変化させ実験を行った epoch 数ごとの動作識別結果を図 4 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 [Lea 16] 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 (数値が高い方が良い結果となる) をそれぞれ求め評価を行った．また，causal モデルでフレームサイズ d を 40 とした場合の 1000 epochs 時の上位 n 件を正解とした場合の結果を図 5 に示し，200 epochs 時，500 epochs 時の 58 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 1 に示す．表 1 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

3.2.3 考察

図 4 より，epoch 数を増やすことで，正解率・編集スコアともに高くなっていくこと，また 1000 epochs の場合，フレームサイズ d は 40 で causal モデルの場合が最も精度が高くなることもわかった．図 5 より，上位 5 位以上を正解とした場合は正解率が 77%以上となる．表 1 より，epoch 数を増やすことで訓練用の動画像数が少ない単語 (例えば，squeeze や peel, take apart など) でも F 値が高くなることがわかる．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

∗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

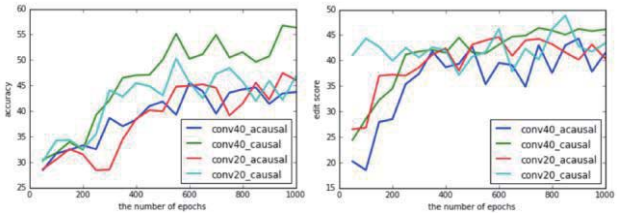

図 4: Accuracy(左) と edit-score(右)

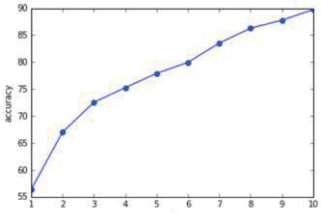

図 5: 上位 10 までの正解率

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 epoch 数を増やしても F 値は上がらなかったと考えられる．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や少ない訓練データ数でも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 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．

3.3.2 実験結果

SSD を実験した結果は図 6 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 6 本で 5 本を訓練用，1 本を評価用に使用した．表 2 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ため，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 5 本の動画像における 112 文と 5,715 枚の画像で学習を行った．

表 1: 単語ごとの比較

単語	200 epochs { 適合率, 再現率, F 値 }	500 epochs { 適合率, 再現率, F 値 }	動画像数
take out	0.45, 0.76, 0.57	0.49, 0.68, 0.55	115
wash	0.14, 0.36, 0.20	0.14, 0.35, 0.20	99
cut	0.55, 0.34, 0.42	0.22, 0.36, 0.27	95
peel	0.06, 0.49, 0.11	0.31, 0.34, 0.32	45
take apart	0.0, 0.0, 0.0	0.17, 0.14, 0.15	37
squeeze	0.0, 0.0, 0.0	0.58, 0.55, 0.56	8
tap	0.0, 0.0, 0.0	0.0, 0.0, 0.0	4

表 2: 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

システムの実装に関しては，言語モデル [Sutskever 14] を深層学習フレームワーク Keras で実装したコード [*6] 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 30 単語を合わせ 63 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 } の 4 次元を入力次元に追加し，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て実験した．

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 6 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 3 に示す．

表 3: 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

3.4.3 考察

表 3 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 6 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，(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，SSD の結果では現れていない単語 ((1)ingredient，(3)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．(5)(6) は評価用の動画像において近いフレーム同士であり，冷蔵庫の隣で食材のパッケージを取っている動画像である．動作識別時の正解ラベルは remove-package としているが訓練データ数が少なかったためにラベルが識別できず，適切に説明文を生成することができなかったと考えられる．

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 を用いた動作を表す単語を選

∗5 https://github.com/rykov8/ssd_keras

∗6 https://github.com/farizrahman4u/seq2seq

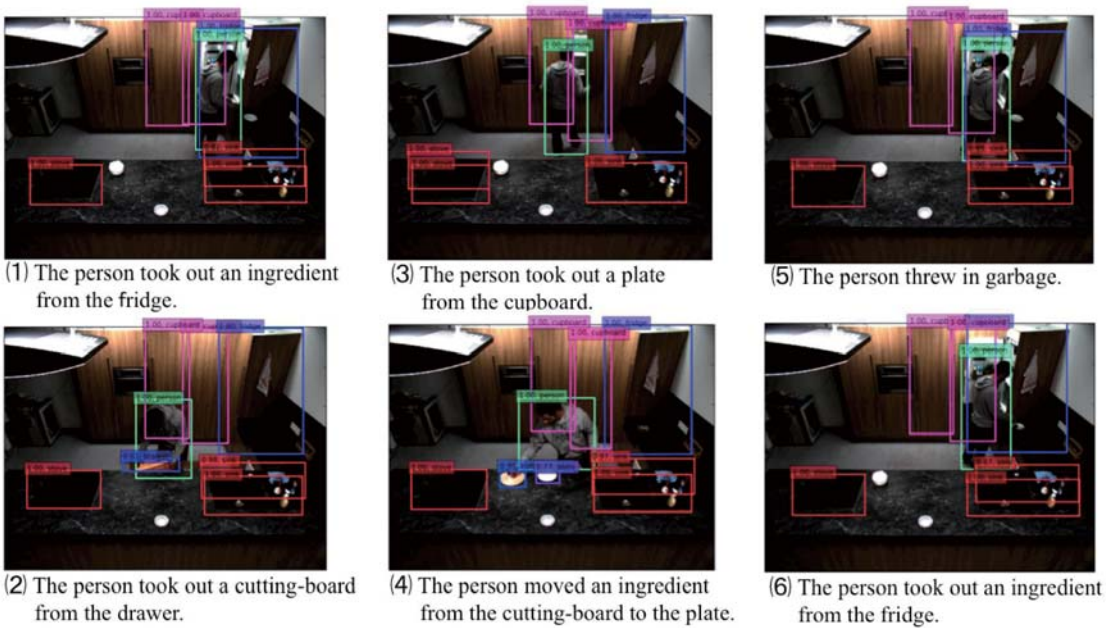

図 6: SSD 結果と生成文の例

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．文生成実験においては人と物体との位置関係を捉えた文が生成されていること，物体検出で検出できない単語も訓練用データから補完できることを確認した．

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や訓練データ数が少な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，画像・動画特徴量から文生成する既存手法との比較より提案手法の有効性を確認し，BLEU スコアが向上するよう改良を進めてきたい．

参考文献

[Donahue 15] Donahue, J., Hendricks, L.A., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., and Darrell, T.: Long-term Recurrent Convolutional Networks for Visual Recognition and Description, In CVPR ' 15 (2015).

[Kiros 15] R. Kiros, R. Salakhutdinov, R. S. Zemel, "Unifying visual-semantic embeddings with multimodal neural language models", In NIPS Deep Learning Workshop, 2015.

[Loffe 15] Ioffe, S. and Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift, In arXiv preprint arXiv:1502.03167 (2015).

[Vinyals 15] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator", In CVPR, 2015.

[Fang 15] 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.

[Pasunuru 17] R. Pasunuru, and M. Bansal, "Multi-Task Video Captioning with Video and Entailment Generation", In ACL, 2017

[Tran 15] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning Spatiotemporal Features with 3D Convolutional Networks", in ICCV, 2015

[Dong 16] J. Dong, X. Li, W. Lan, Y. Huo, C. G. M. Snoek, "Early Embedding and Late Reranking for Video Captioning", In ACMMM, 2016

[Cao 17] Z. Cao, T. Simon, S. Wei, and Y. Sheikh."Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields", In CVPR, 2017.

[Lea 16] 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.

[Liu 16] 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.

[Ren 16] S. Ren, K. He, R. Girshick, and J. Sun. 2016. Faster r-cnn: Towards real-time object detection with region proposal networks. arXiv preprint arXiv:1506.01497.

[Redmon 16] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. 2016. You only look once: Unified, real-time object detection. arXiv preprint arXiv:1506.02640.

[Sutskever 14] I.Sutskever, O. Vinyals, and Q. V. Le. "Sequence to sequence learning with neural networks", In Advances in NIPS, 2014.

[Regneri 13] 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.

[Senina 14] 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.