강화학습을 통한 계층적 RNN의 행동 인식 성능강화

김상조[O], 곽소항[*], 차의영[*]
[O*]부산대학교 전기전자컴퓨터공학과
e-mail: kimsjpk@hanmail.net, eycha@pusan.ac.kr

Improved the action recognition performance of hierarchical RNNs through reinforcement learning

Sang-Jo Kim[O], Shao-Heng Kuo[*], Eui-Young Cha[*]
[O*]Dept. of Electricity and Electronic Computer Science Engineering, Pusan National University

● 요 약 ●

본 논문에서는 계층적 RNN의 성능 향상을 위하여 강화학습을 통한 계층적 RNN 내 파라미터를 효율적으로 찾는 방법을 제안한다. 계층적 RNN 내 임의의 파라미터에서 학습을 진행하고 얻는 분류 정확도를 보상으로 하여 간소화된 강화학습 네트워크에서 보상을 최대화하도록 강화학습 내부 파라미터를 수정한다. 기존의 강화학습을 통한 내부 구조를 찾는 네트워크는 많은 자원과 시간을 소모하므로 이를 해결하기 위해 간소화된 강화학습 구조를 적용하였고 이를 통해 적은 컴퓨터 자원에서 학습속도를 증가시킬 수 있었다. 간소화된 강화학습을 통해 계층적 RNN의 파라미터를 수정하고 이를 행동 인식 데이터 세트에 적용한 결과 기존 알고리즘 대비 높은 성능을 얻을 수 있었다.

키워드: 계층적 RNN(Hierarchical RNN), 강화학습(reinforcement learning), 행동 인식(action recognition)

I. Introduction

사회 안전 증대를 위하여 도로와 공공장소 주변에 많은 CCTV 감시카메라를 구축하였으나 이를 분석하는데 필요한 인력의 부족으로 인한 관리 소홀의 취약점을 지니고 있다. 이를 보완하는 방법으로 딥러닝을 적용한 행동인식을 사용하여 자능형 감시스템을 구축함으로써 사회 안전망 구축에 기여할 수 있다. 행동인식에 적용되는 딥러닝 알고리즘으로 3D 컨볼루션을 이용하는 방법[1]과 depth map, skeleton을 적용한 방법[2] 등의 연구가 이루어지고 있다. 3D 컨볼루션의 경우 처리 데이터양이 많으므로 학습을 위해 높은 하드웨어 사양이 요구되며 depth map, 스켈레톤 정보를 얻기 위하여 kinect와 같은 추가적인 장비가 요구된다. 본 논문에서는 이를 극복한 방법을 제안한다. 그리고 NAS(Network Architecture Search)[3]의 경우 데이터 세트에 적합한 딥러닝 네트워크의 구조 및 모델 파라미터를 찾아주어 분류성능 향상을 가능하게 하나 이를 위하여 많은 하드웨어 사양과 시간이 소모되는 단점을 지니고 있다. 본 논문에서는 한정된 파라미터만을 대상으로 하는 간소화된 NAS를 적용하므로 NAS의 단점을 극복하였다. 간소화된 NAS를 기존의 계층적 RNN을 사용한 행동 인식 알고리즘에 적용하여 데이터 세트에 맞는 계층적 RNN의 파라미터를 수정한 모델을 행동 인식 데이터세트에 적용함으로써 행동 인식 성능이 향상함을 확인하였다.

II. Preliminaries

1. SYSU 3D Human-Object Interaction Set

위 데이터 세트[4]는 40명이 각각 12가지 행동(전화 받기, 가방 메기 등)에 대하여 자유롭게 행동을 하고 그 영상을 저장하였다. 데이터 세트는 RGB 영상과 깊이 정보, 스켈레톤 정보를 포함하나 본 논문에서는 RGB 영상만을 사용하여 실험을 진행하였다.

2. 스켈레톤 추출 알고리즘

스켈레톤을 추출하기 위해 ArtTrack[5]을 사용하여 영상에서 사람의 각 부위 및 관절 포인트를 추출하였다.

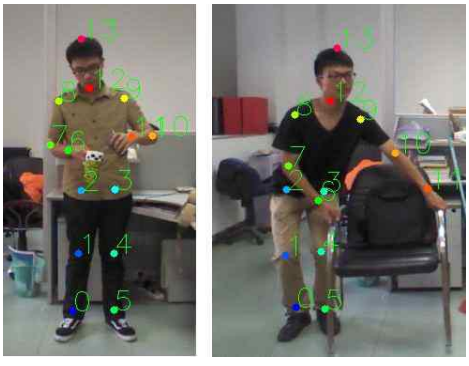

Fig. 1. 데이터세트 내 영상의 포인트 추출 예시

3. 특징 벡터 추출

Du et al이 제안한 행동 인식 알고리즘[6]의 방법을 적용하고 수식 1을 사용하여 신체 별 포인트의 벡터를 얻고 벡터의 x, y 좌표를 묶어 Fig. 2와 같이 다섯 부분으로 나누어 계층적 RNN 구조의 입력데이터로 사용하였다. 식에서 p는 포인트의 x, y 좌표, i는 포인트의 순서, s는 각 프레임의 순서를 나타낸다.

$$v_i^s = \frac{p_i^{s+1} - p_i^s}{\triangle T} | 1 < s < \tau$$

$$f(x,y) = v_i^s(x,y)$$

수식 1. 벡터 관계 식

각 프레임 마다 사람의 스켈레톤 정보에서 부위 별 벡터로 변환하고 연속된 50개의 프레임별 벡터를 묶어 시간 연속된 데이터를 만든다. 데이터 세트의 크기가 480개의 avi로 구성되어 있어서 데이터 증대를 위하여 프레임 내 40개의 인덱스를 추출하고 이후의 50개 프레임의 시간 연속된 데이터를 학습 데이터로 사용하였다. 그리고 training, validation, test data를 각각 4:1:1(12800:3200:3200)로 구성하였다.

4. 계층적 RNN 구조

Du et al이 제안한 행동 인식 알고리즘[6]의 계층적 RNN의 구조를 사용하였다. 그리고 성능향상을 위하여 계층 내 각 RNN은 Nested LSTM(Long Stort- Term Memory)[7]을 사용하였다.

그리고 fully connected layer의 활성화 함수로 SELU[8]를 사용하였다.

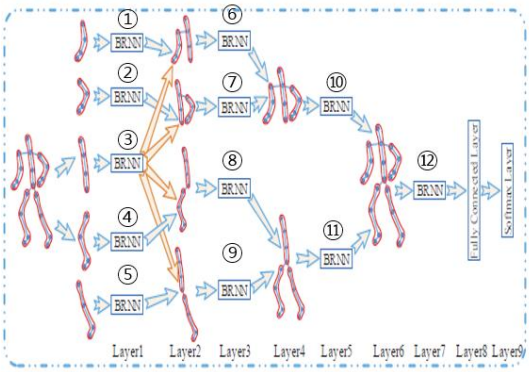

Fig. 2. 본 논문에서 사용된 계층적 RNN 구조 그림

5. 간소화된 NAS

데이터 세트에 딥러닝 적용 시 데이터세트에 맞는 신경망을 설계하고 각 파라미터를 적용하는데 드는 시간과 노력을 줄일 수 있도록 강화학습을 통하여 자동으로 찾을 방법으로 NAS[3]가 제안되었다. 본 논문의 계층적 RNN과 같은 네트워크에서 학습을 통해 얻은 분류 정확도를 보상으로 하여 NAS 네트워크 내 파라미터를 조정하고 분류 정확도의 변화를 얻는 policy gradient 방법을 적용하고 이를 반복 시행하고 최적의 파라미터 및 네트워크 구조를 자동으로 찾게 된다. NAS에서는 활성화 함수와 RNN 네트워크를 바꾸는 구조로 되어 있으나 학습을 위하여 많은 자원과 시간이 소모되므로 본 논문에서는 Nested LSTM의 입력 노드와 depth의 개수를 조정하는 간소화된 NAS를 적용하였다. NAS에서의 state는 입력 데이터에 맞게 4(1배), 6(1.5배), 8(2배)의 노드와 Nested LSTM의 depth로 1, 2를 사용하여 Fig. 2에서의 12개의 RNN에 적용하였다. 이에 간소화된 NAS의 총 state 수는 $12,288\,(3*2^{12})$이고 본 논문에서는 200번의 탐색을 통하여 결과를 얻고 그중 가장 높은 분류 정확도의 파라미터를 행동 인식 알고리즘에 사용하였다. 가장 높은 분류 정확도의 파라미터는 Fig. 2의 layer 3 내 8번째 Nested LSTM의 depth는 2이고 나머지는 1이며, Layer 1의 ①~⑤의 node는 6, 6, 3, 6, 6(1.5배)이다. Nested LSTM의 모든 layer의 depth를 1로 했을 때와 2로 할 때, 그리고 수정된 파라미터를 한 계층적 RNN 모델에서 epoch을 20으로 두고 분류 정확도 비교 시 수정된 파라미터에서의 정확도가 높음으로 보아 강화학습을 적용한 모델의 성능이 향상됨을 확인하였다.

Table 1. SYSU 3D HOI 데이터세트에 관한 파라미터 간 성능 비교

Nested LSTM	정확도(%)
depth = 1, node 1배	56.06
depth = 2, node 1배	8.25
강화학습 후 수정된 파라미터	58.03

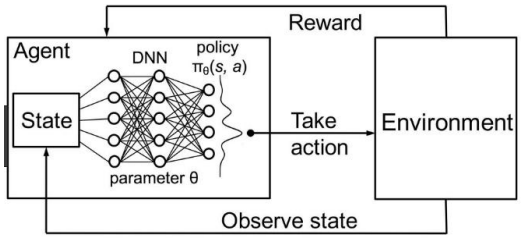

Fig. 3. 강화학습 구조 예시

III. Experiment

SYSU 3D Human-Object Interaction Set 영상의 RGB 데이터에서 ArtTrack[5]을 사용하여 스켈레톤 정보를 추출하고 이를 Fig. 2와 같이 5개의 부분 벡터로 변형하였다. 샘플링된 인덱스와 이후 50 프레임을 묶고 간소화된 NAS를 적용하고 계층적 RNN 파라미터를 수정한 모델을 사용하여 epoch을 100으로 두고 실험을 진행하였다. 그리고 표 1과 같이 기존 알고리즘 대비 높은 행동 인식 분류 정확도를 얻을 수 있었다.

Table 2. SYSU 3D HOI 데이터세트에 관한 알고리즘 간 성능 비교

방법	정확도(%)
HON4D[9]	73.39
HFM[10]	75.03
MPCCA[11]	76.25
SVM[12]	77.34
MTDA[13]	79.19
JOULE[12]	79.63
제안한 방법	97.00

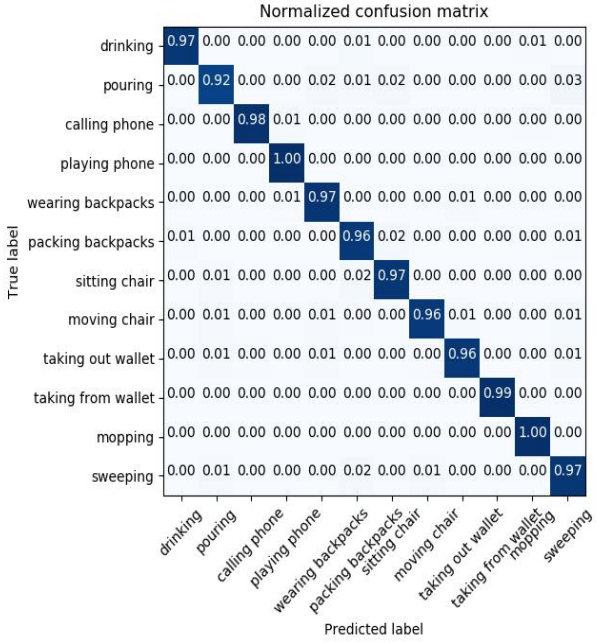

Fig. 4. SYSU 3D HOI 데이터세트에 대한 confusion matrix

IV. Conclusions

본 논문에서는 RGB 영상 입력만을 사용하여 행동 인식 분류 알고리즘을 제안하였다. 학습의 성능을 높이려는 방법으로 샘플링을 통한 데이터 증대, SELU, 강화학습을 통한 모델 내 파라미터 변형 방법을 사용하고 데이터 세트의 테스트데이터에 적용하여 높은 행동 인식 분류 정확도를 얻을 수 있었다. 그리고 실험결과를 바탕으로 본 논문에서 제안한 방법이 다른 계층적 RNN의 성능향상에 기여할 수 있을 것으로 기대된다.

REFERENCES

[1] TRAN, Du, et al. Learning spatiotemporal features with 3d convolutional networks. arXiv preprint arXiv:1412.0767, 2014.

[2] LI, Wanqing; ZHANG, Zhengyou; LIU, Zicheng. Action recognition based on a bag of 3d points. In: Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on. p. 9-14. IEEE, 2010.

[3] ZOPH, Barret; LE, Quoc V. Neural architecture search with reinforcement learning. arXiv preprint arXiv:1611.01578, 2016.

[4] HU, Jian-Fang, et al. Jointly learning heterogeneous features for RGB-D activity recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. p. 5344-5352. 2015.

[5] INSAFUTDINOV, Eldar, et al. ArtTrack: Articulated multi-person tracking in the wild. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2017.

[6] Du, Yong, Wei Wang, and Liang Wang. "Hierarchical recurrent neural network for skeleton based action recognition." Proceedings of the IEEE conference on computer vision and pattern recognition. 2015.

[7] MONIZ, Joel Ruben Antony; KRUEGER, David. Nested LSTMs. arXiv preprint arXiv:1801.10308, 2018.

[8] KLAMBAUER, Günter, et al. Self-Normalizing Neural Networks. arXiv preprint arXiv:1706.02515, 2017.

[9] WANG, Jiang, et al. Robust 3d action recognition with random occupancy patterns. In: Computer vision–ECCV 2012. Springer, Berlin, Heidelberg, p. 872-885. 2012.

[10] CHAARAOUI, Alexandros; PADILLA-LOPEZ, Jose; FLÓREZ-REVUELTA, Francisco. Fusion of skeletal and silhouette-based features for human action recognition with rgb-d devices. In: Proceedings of the IEEE international conference on computer vision workshops. p. 91-97. 2013.

[11] CAO, Liangliang, et al. Heterogeneous feature machines for visual recognition. In: Computer Vision, 2009 IEEE 12th International Conference on. IEEE, p. 1095-1102. 2009.

[12] HU, Jian-Fang, et al. Jointly learning heterogeneous features for RGB-D activity recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. p. 5344-5352. 2015.

[13] ZHANG, Yu; YEUNG, Dit Yan. Multi-task learning in heterogeneous feature spaces. In: 25th AAAI Conference on Artificial Intelligence and the 23rd Innovative Applications of Artificial Intelligence Conference, AAAI-11/IAAI-11, San Francisco, CA, United States,

7-11 August 2011, Code 87049, Proceedings of the National Conference on Artificial Intelligence. p.574. 2011.