RGB 데이터 기반 행동 인식에 관한 연구

김상조*, 김미경*, 차의영*
*부산대학교 전기전자컴퓨터공학과
e-mail:kimsjpk@naver.com

A Study on Action Recognition based on RGB data

Sang-Jo Kim*, Mi-Kyoung Kim*, Eui-Young Cha*
*Dept of Electricity and Electronic Computer Engineering, Pusan National University

요 약

최근 딥러닝을 통하여 영상의 카테고리 분류를 응용한 행동 인식이 활발히 연구되고 있다. 그러나 행동 인식을 위한 기존 연구 방법은 높은 수준의 하드웨어 사양을 요구하며 행동 인식에 대한 학습에 많은 시간이 소모되는 문제점을 지니고 있다. 또한, 행동 인식 테스트 결과를 얻기 위해 많은 시간이 소모되며 딥러닝 특성상 적은 수의 학습 데이터는 overfitting 문제를 일으킨다. 본 연구에서는 이러한 문제점을 해결하고자 행동인식을 위한 학습시간과 테스트 시간 감소를 위해 미리 학습된 VGG 모델을 사용해 얻어낸 RGB 데이터의 특징만을 학습에 사용하고 적은 수의 데이터로 행동 인식 테스트 결과를 높이기 위하여 RGB 데이터 증대를 통해 기존의 행동인식 연구보다 학습시간과 행동인식 테스트에 소모되는 시간을 줄인 방법을 행동 인식에 적용하였다. 이 방법을 UCF50 Dataset 에 적용하여 98.13% 의 행동인식에 관한 정확성을 확인하였다

1. 서론

한 비디오 기반의 행동 인식에 관한 연구는 안전 분야 및 행동 분석을 통한 애플리케이션 개발을 위해 학계에 커다란 관심을 불러일으키며 활발히 연구 중이다[1]. 크게 두 가지 방향으로 연구가 진행 중인데 첫째가 Bag of Visual Words(BoVWs)와 같은 직접 설계를 통하여 특징을 추출하는 방법[2]이고 두 번째는 convolutional network(ConvNet)를 사용하여 비디오의 데이터를 사용하여 행동 인식에 관한 연구가 진행되고 있다. 현재 convolutional network를 사용하여 방법 중 RGB 데이터와 optical flow를 사용한 two-stream ConvNets를 사용한 방법을 통하여 활발히 연구가 진행[1] 중인데 이 방법을 행동 인식에 적용하면 두 모델을 메모리상에 로드하기 때문에 큰 용량의 비디오 램이 필요하며 RGB 데이터와 optical flow 데이터를 모두 사용하여 동영상의 행동 인식 분석에 사용하기 때문에 행동 인식을 테스트하는데 많은 시간이 걸리는 단점을 지니고 있다. 그리고 행동 인식을 위한 학습데이터는 ImageNet Dataset[3] 과 비교했을 때 상대적으로 아주 적은 dataset을 지녀 overfitting 문제를 지니고 있다. 위의 문제를 해결하기 위해 1) RGB 데이터만을 미리 학습된 CNN 모델인 VGGNet[4]을 사용하여 RGB 데이터의 특징을 뽑아내어 행동인식을 위한 학습데이터로 이용하고 2) Drop-out 사용하고 3) 학습 데이터 부족에 의한 overfitting을 피하고자 학습 데이터를 증가 방법으로 multi-scale cropping을 이용하여 ConvNet의 학습에 사용하였다. 이를 통해 2가지 모델을 적용한 행동 인식보다 학습시간과 행동 인식 테스트에 걸리는 시간을 줄일 수 있었고 UCF50 Dataset 에 적용한 결과 98.13%의 행동 인식 정확도를 확인하였다.

2. 본문

2.1 네트워크 모델

ConvNet에 사용된 미리 학습된 모델(VGGNet) 사용

행동 인식에 사용된 Dataset 은 상대적으로 작으므로 학습 시 overfitting 문제에 직면하게 된다. 이를 위해 미리 학습된 모델의 초깃값을 사용하여 ConvNet를 학습 시 효과적인 학습을 진행할 수 있다.[1]

Drop-out 사용

다른 ConvNet을 사용하여 행동 인식을 한 방법[7]과 같이 0.9의 drop-out 방법을 fully connected layer에 적용하여 overfitting 문제 해결에 적용하였다.

데이터 증대 기법 사용

데이터 증대 기법 중 random cropping과 multi-scale

cropping을 사용하여 데이터를 증대시켰다. 256 x 340의 입력 데이터를 256, 224, 192, 168 사이즈로 자른 후 crop region을 224 x 224로 크기조정 후 학습데이터로 사용하였다.

네트워크 테스트 방법

입력 데이터 동영상 파일의 frame 중 25 frame을 샘플데이터로 사용하여 테스트를 진행하였다. 샘플데이터와 샘플데이터를 cropping 한 각 이미지의 prediction score 평균값을 계산 후 행동 인식 분류를 진행하였다. 실험은 Caffe framework[5]를 통해 진행하였고 사양은 I-7 6770, Nvidia GTX 1080을 사용하였다.

2.2 실험

(그림 1). UCF50 dataset 비디오의 스크린샷

Dataset(UCF50)

Youtube에 있는 50가지 실생활 action categories로 이루어져 있는 dataset으로 각 행동 class 당 최소 100개 이상의 video로 총 6,676개의 video로 구성되어 있다.[6]

2.3 결과

미리 학습된 VGGNet-16을 10000번의 iteration 후 training data : validation data : test data를 3 : 1 : 1로 나누어 행동 인식에 대한 학습과 테스트를 진행하였고 1,337개의 임의의 테스트 데이터에 대하여 98.13%의 정확도를 확인할 수 있었다(1312 / 1337)

3. 결론

이 논문에서는 행동 인식을 위한 입력데이터의 수가 작아 overfitting의 문제를 해결하기 위해 Drop-out과 데이터 증대 기법을 사용하고 기존의 행동인식을 위한 높은 사양의 하드웨어 요구사항을 극복하고 빠른 학습(2 hour/ 10000 iterations)과 테스트(140sec/ 80 videos) 결과를 얻기 위해 기존의 two stream 방법이 아닌 RGB 데이터만을 ConvNet의 입력데이터로 사용하여 행동 인식을 진행하였다. 위의 방법을 UCF50 Dataset 에 적용하여 98.13%의 정확도를 확인할 수 있었다.

참고문헌

[1] Simonyan, K., Zisserman, “A.: Two-stream convolutional networks for action recognition in videos”. In: NIPS, pp. 568 - 576, 2014.

[2] H. Wang and C. Schmid. “Action recognition with improved trajectories“. In ICCV, pages 3551 - 3558, 2013. 1,4

[3] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and F. Li. “ImageNet: A large-scale hierarchical image database.” In CVPR, pages 248 - 255, 2009. 1

[4] K. Simonyan and A. Zisserman. “Very deep convolutional networks for large-scale image recognition“ CoRR, abs/1409.1556, 2014. 1, 2, 3

[5] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. “Caffe: Convolutional Architecture for Fast Feature Embedding”. In Proceedings of the 22nd ACM international conference on Multimedia (MM '14), 2014

[6] Reddy, K.K. & Shah, M. “Recognizing 50 human action categories of web videos”, Machine Vision and Applications 2013. 24: 971.