Video Action Classification 최신 기술 조사

차진혁, 정승원
동국대학교 멀티미디어공학과
e-mail: ckwlsgur20@google.com

A Survey on Recent Video Action Classification Techniques

Jin Hyuck Cha, Seung-Won Jung
Dept. of Multimedia Engineering, Dongguk University

요 약

최근 딥러닝을 이용해 정지 영상에 대한 연구 뿐만 아니라 동영상에 대한 연구들이 진행되고 있다. 본 논문에서는 동영상 딥러닝 기술에서 가장 주가 되고 있는 video action classification 에 대한 최신 기술들을 조사했다.

1. 서론

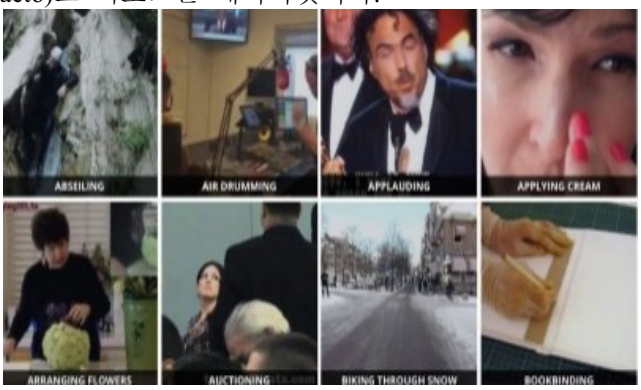
최근 딥러닝을 이용해 이미지에 대한 연구 뿐만 아니라 시간(temporal) 정보를 가진 video 에 대한 연구들이 진행되고 있다. 대표적으로 video 에 대해 설명을 자동으로 달아주는 video captioning[1], video 에서 불필요한 부분을 지워주는 video inpainting[2] 같은 분야에서 활발히 연구되고 있다. 특히나 video 분야에서 가장 주가 되는 분야는 video action classification 이다. 해당 video 에 있는 사람의 action 을 인식하고, 이에 대해 분류를 하는 분야이다.

2. Video Action Classification Dataset

Video action classification 에서 자주 쓰이는 데이터셋에 대해 소개한다. 첫번째는 UCF-101[3] 데이터셋이다. 13,320 개의 video 들과 101 개의 action class 를 가졌다. 다양한 카메라 움직임과 배경, 조명을 가진다. 일반적으로 test benchmark 로 쓰인다.

<그림 1. UCF-101 데이터셋>

두번째는 Kinetic[4] 데이터셋이다. 딥마인드에서 만든 데이터셋으로 각 action class 에 맞는 비디오의 유튜브 주소가 들어 있다. 총 650,000 개의 비디오와 700 개의 action class 로 이루어져 있다. 각 비디오는 10 초 가량으로 이뤄져 있으며, 하나의 class 레이블이 달려있다. 각 비디오의 레이블링은 사람이 했다. 최근 action classification 의 학습 데이터셋으로 산업계 표준(de facto)로 떠오르는 데이터셋이다.



<그림 2. Kinetics 데이터셋>

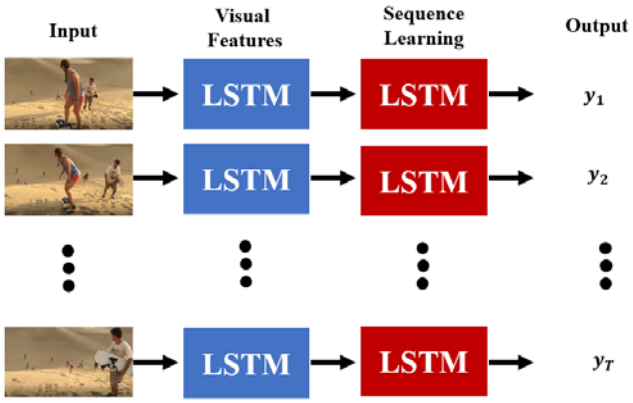
3. Action Classification Networks

Video action classification 분야에서 중추가 되는 모델들에 대해 소개한다.

3.1 Long-term Recurrent Convolutional Networks

LRCN(Long-term Recurrent Convolutional Networks)[5]는 딥러닝을 video 에 적용한 초창기의 구조이다. encoder-decoder 구조를 지녔으며, encoder 에는 2D CNN(Convolution Neural Networks)를 사용, decoder 에는 RNN(Recurrent Neural Network)를 사용했다. 모델의 입력으로는 RGB 이미지혹은 optical flow 를 넣어주어 action class 를 예측한다. optical flow 를 넣어서 나온 예측 값과 RGB 이미지를 넣어 나온 예측 값의 평균으로 예측한 결과가 성능이 가장 좋았다. 이 모델의 주요 기여는 video action classification 분야에서 RNN 기반의 구조를 제안했으며, encoder-decoder 구조를 제안,

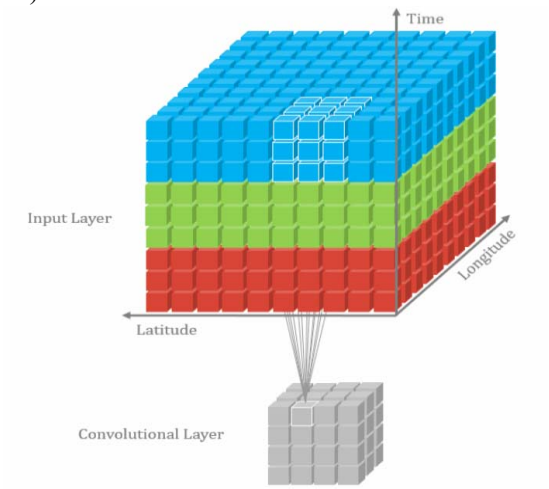
end-to-end 학습 방식을 제안했다. 하지만 이 방식에는 LSTM 의 고질적인 문제인 vanishing gradient 로 인해 긴 시간 정보(long range temporal information)를 학습하기가 힘들다는 점이 있다.



<그림 3. LRCN>

3.2 3D Convolution Neural Networks

3D CNN(3D Convolution Neural Network)[6]는 기존의 2D CNN 이 시공간(spatio-temporal) 정보를 잘 파악하지 못한다는 특성을 보완하기 위해 제안된 구조이다. 2D convolution 은 출력이 2D 이지만, 3D Convolution 은 출력이 3D 형태로 나온다. (그림 5. 참조)

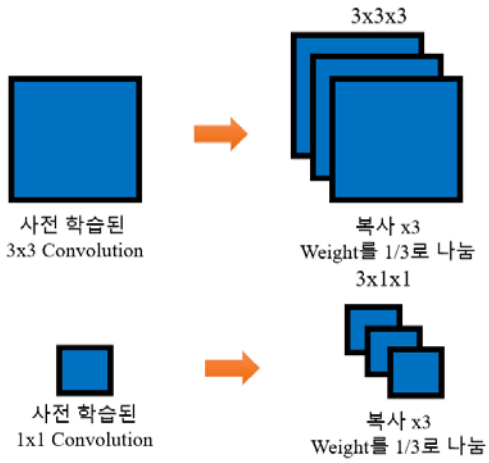


<그림 5. 3D Convolution>

주요 기여로는 feature 추출기로서 3D convolution 을 제안, 3D convolution 의 커널 크기와 모델에 대한 전반적인 조사, deconvolution layer 를 모델 결정에 대한 해석으로서 사용했다는 점이다. 하지만 여전히 긴 시간 정보를 잘 학습하지 못한다는 문제가 남아있었고, 계산량의 문제가 굉장히 크다.

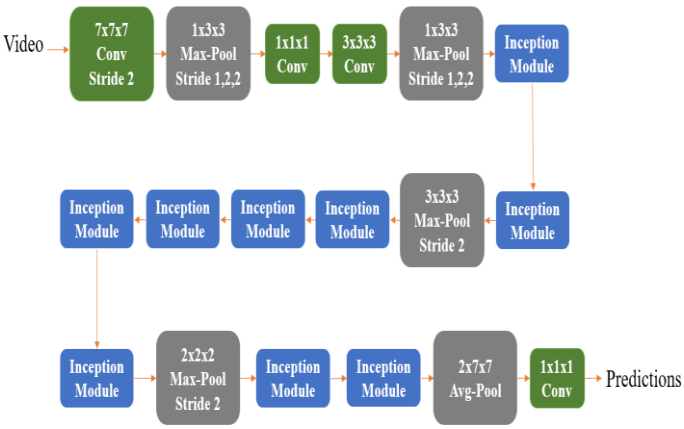
3.3 Inflated 3D Convolutional Neural Networks

I3D(Inflated 3D Convolutional Neural Networks)[7]는 video action classification 에서 굉장히 중요한 논문이다. 해당 논문에서 kinetics 데이터셋을 만들고 이를 학습 데이터셋으로 사용해 I3D 모델을 학습시켰다. I3D 모델의 핵심은 ImageNet[8] 데이터셋으로 사전 학습(pretrained)된 2D CNN 을 3D CNN 으로 바꾸는 것이다. 저자는 이러한 NxN 필터를 NxNxN 필터로 바꾸는 방식을 'inflating'이라고 명명했다. 필터의 차원을 늘려준 후, 늘려준 만큼 weight 를 1/N 해준다. 예를 들어 3x3 필터의 weight 를 1/3 해주고, 3x3 필터를 3 개 복제해 3x3x3 필터로 만들어준다. (그림 6. 참조)



<그림 6. Inflating>

논문에서 제안한 모델 구조는 inception module[9]을 활용한 구조이며, optical flow 를 입력으로 넣어 학습을 하는 모델과, RGB 이미지를 입력으로 넣어 학습을 하는 모델 두개로 구성했다. 두 모델의 예측 결과의 평균을 내어 최종 예측을 한다.



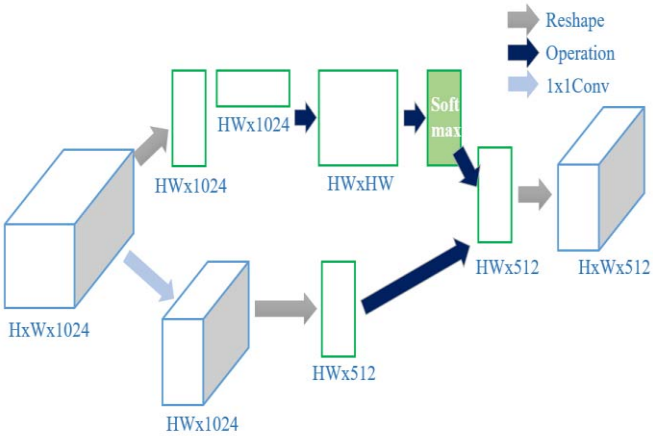
<그림 7. I3D>

주여 기여로는 video action classification 분야에서도 ImageNet 에 버금가는 대규모 데이터셋을 공개했다는 점, inflating 이라는 참신한 기법을 이용해 큰 성능 변화를 이끌어 냈다는 점이다. 특히나 I3D 는 발표 이후 대다수의 video action classification 모델의 중추가 되었다.

3.4 Non-local Neural Networks

Non-local Neural Networks[10]는 non-local block 이라는 아주 간단하면서도 활용도가 무궁무진한 구조를 제안했다. Non-local Neural Network 는 기존의 convolution 이

가지고 있던 문제점인 지역적인 영역만을 본다는 문제를 해결하였다. 즉 receptive field 크기 관점에서 보면, 지역적 연산을 여러 차례 쌓으면 receptive field 의 크기를 키울 수 있지만, 아무리 많이 쌓더라도 한번에 전체 영역을 살펴 볼 수는 없다. 하지만 저자가 제안한 non-local block 을 사용하면, 전체 영역을 볼 수 있게 되어 지역적인 단점을 해결할 수 있다. Non-local block 은 들어온 feature 에 대해 similarity 연산을 해 feature 재조정을 한다고 주장한다.(그림 8 참조)

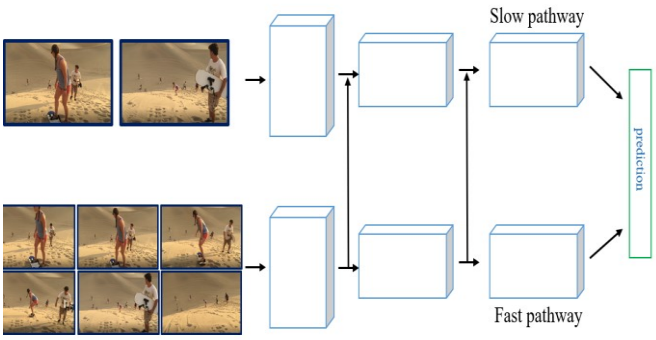


<그림 8. Non-local block>

주요 기여로는 구조를 크게 변경하지 않고 모델의 맨 마지막에 non-local block 을 2 개 넣어준 것만으로도 성능상으로 2~4%가 증가한다. 또 비디오 영역 뿐만 아니라 이미지 분류에서도 non-local block 을 넣어준 것만으로 성능 향상을 보여줬다.

3.5 SlowFast Networks

SlowFast Networks[11]는 사람 눈의 세포에 따른 물체 정보 인식에 관한 연구에 착안해 제안된 모델 구조이다. 기존에 주로 쓰이던 모델들은 입력으로 RGB 이미지와 optical flow 를 쓰며, 각각의 모델을 따로 학습시켰다. SlowFast Networks 에서는 RGB 이미지만을 넣어주어 end-to-end 방식으로 학습을 한다. 논문에 따르면 사람의 눈은 P-cells 과 M-cells 로 이뤄져 있는데, 그 중 M-cells 는 빠른 시간적 주기(high temporal frequency)에 대한 정보를 받아들이고, 빠른 시간적 변화(fast temporal change)에 반응한다. 그러나 공간적 자세한 특성(spatial detail)에는 거의 반응하지 않는다. 반면 P-cells 는 공간적 자세한 특성에 대해 반응하지만, 시간적 정보에는 반응하지 않는다는 특징을 지닌다. 이런 특징을 딥러닝 모델에 적용해 RGB 이미지에 대해 서로 다른 샘플링 비율을 정해 두개의 모델을 학습시킨 후 최종적으로 feature 를 합쳐 예측을 하는 모델을 제안했다. 샘플링 비율을 높게 잡아 입력 비디오의 변환이 빠른 모델은 “slow pathway”라 부르고, 샘플링 비율을 낮게 잡아 입력 비디오의 변환이 느린 모델은 “fast pathway”라 부른다. (그림 9. 참조)



<그림 9. SlowFast Network>

주요 기여로는 모델을 사람의 인지 시스템을 모방해 만들었고, 논문에서 다양한 실험을 통해 성능을 증명했다. 특히나 기존에 성능상으로 대세였던 optical flow 와 RGB 이미지를 둘 다 쓰는 방식에서 벗어나, optical flow 를 안쓰기에, 성능과 연산량 측면 모두에서 기존 모델들을 압도하였다. 앞으로 video action classification 분야에서 중추가 될 가능성이 높은 모델이다.

4 결론

Video action classification 은 입력 영상의 차원이 커 연산량이 너무 많고, 대규모의 데이터셋이 많이 존재하지 않아 2D 이미지 분류보다 발전이 더딘 분야였다. 하지만 2017 년 kinetics 데이터셋의 공개와 함께 많은 연구 팀들이 지속적으로 논문을 발표하였고, 최근 나온 SlowFast Networks 에서 좋은 성능을 지닌 모델들이 연구되며 빠르게 발전하고 있다. 하나 주목할 만한 점은 모델의 입력으로 오는 영상에 대해 optical flow 와 같이 사람이 만든 feature 를 넣는 방식은 도태되고 RGB 이미지만을 넣어, 모델이 알아서 feature 를 추출하도록 하도록 발전하고 있다는 점이다. 이 추세가 계속된다면 optical flow 와 같이 사람이 설계한 (hand-crafted) feature 추출 방식을 모델에 입력으로 넣는 방식은 뒤쳐질 것이라 예상된다.

참고문헌

[1] Venugopalan, et al. "Sequence to sequence-video to text." *Proceedings of the IEEE international conference on computer vision*. 2015.

[2] Kim, Dahun, et al. "Deep Video Inpainting." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2019.

[3] Soomro, et al. "UCF101: A dataset of 101 human actions classes from videos in the wild." *arXiv preprint arXiv:1212.0402*.

[4] Kay, et al. "The kinetics human action video dataset." *arXiv preprint arXiv:1705.06950*.

[5] Donahue, et al. "Long-term recurrent convolutional networks for visual recognition and description." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015.

[6] Ji, et al. "3D convolutional neural networks for human action recognition." *IEEE transactions on pattern*

analysis and machine intelligence 35.1: 221-231.
[7] Carreira, et al. "Quo vadis, action recognition? a new model and the kinetics dataset." *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017.
[8] Deng, et al. "Imagenet: A large-scale hierarchical image database." *2009 IEEE conference on computer vision and pattern recognition*.
[9] Szegedy, et al. "Going deeper with convolutions." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015.
[10] Wang, et al. "Non-local neural networks." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018.
[11] Feichtenhofer, et al. "Slowfast networks for video recognition." *arXiv preprint arXiv:1812.03982*