Convolutional Neural Network를 이용한 동영상 내 동작 인식을 위한 효과적 학습 방법

주현진 ○ 낭종호

서강대학교 컴퓨터공학과

hjjoo123@sogang.ac.kr, jhnang@sogang.ac.kr

An Effective Training Method for Action Recognition in Videos by Using Convolutional Neural Network

Hyunjin Joo○ Jongho Nang

Department of Computer Science and Engineering, Sogang University

요 약

최근 딥러닝을 이용한 멀티미디어 분석에 대한 연구가 활발히 진행되면서 다양한 네트워크 모델이 소개되고 있다. 그 중에서도 Convolutional Neural Network (CNN)은 영상에서 고차원의 Feature를 추출하여 인식하는 데 상대적으로 탁월한 성능을 보이면서, 다양한 모델이 계속적으로 연구 및 발표되고 있는 상황이다. 본 논문에서는 대표적인 CNN 구조 중 하나인 AlexNet을 이용하여 UCF-101 데이터 셋을 효과적으로 학습하는 방법을 제안하고, 영상 내 동작 인식 성능을 분석해보고자 한다.

1. 서 론[1]

영상은 음성이나 텍스트에 비해 정보량이 많은 것이 특징이다. 따라서 하나의 벡터를 input으로 하여 hidden layer와 full connection을 이루는 일반적인 Neural Net은 이미지 데이터를 학습하는 데 scalability 측면에서 취약하며, overfitting 문제를 야기할 수 있다.

이에 반해 Convolutional Neural Net (CNN)은 input 이미지에 일정 크기의 필터를 통해 convolution 연산을 적용하는 방식으로 이웃한 layer 사이에서 local connection을 이룬다. 필터의 사이즈가 곧 뉴런의 receptive field를 결정하는 것이다. CNN의 또 다른 특징은 한 layer 내의 뉴런들이 서로 파라미터를 공유한다는 점인데, 이는 결국 layer input에 적용되는 필터가 2차원 sliding window이자 네트워크에서 학습할 weight set이라는 의미이다. 따라서 필터를 어떻게 설정 하느냐에 따라 layer output의 크기 및 특성이 결정된다. 이 때 하나의 필터는 하나의 특성을 반영하므로 여러 개의 필터를 둠으로써 CNN의 특징 중 하나인 뉴런의 3차원 배치가 구성된다. 이렇게 각 필터를 통해 만들어진 output을 Feature map이라 하며, 일반적으로 여러 개의 Convolution layer와 Pooling layer의 set을 지나면서 feature가 추상화됨에 따라 Feature map의 수도 늘어난다.

본 연구는 미래창조과학부 및 정보통신기술진흥센터의 정보통신·방송 연구개발 사업의 일환으로 하였음. [R0126-16-1112, 퍼스널 미디어가 연결공유결합하여 재구성 가능케 하는 복합 모달리티 기반 미디어 응용 프레임워크 개발]

이러한 CNN의 구조는 기존의 Neural Net에 비해 파라미터의 수를 크게 줄였으며, 이미지의 특성을 학습하는 데에도 효과적이었다. 대표적으로 ImageNet 데이터 셋을 이용한 Large Scale Visual Recognition Challenge (ILSVRC)에서 발표된 AlexNet[3], VGGNet[4], GoogLeNet[5], ResNet[6] 등이 있으며, ILSVRC 2015에서 우승한 ResNet의 경우 Residual function의 개념을 도입하여 3.57%의 Top-5 error를 보였다.

본 논문에서는 CNN을 이용한 여러 응용 중에서도 영상 내 동작 인식에 초점을 맞추어, UCF-101 데이터 셋[10]의 101가지의 동작을 학습하고 모델의 인식 성능을 비교 및 분석한다.

2. 관련 연구

동작 인식(Action recognition)이란 기본적으로 사람을 대상으로 하며, 영상에 등장하는 사람의 움직임이나 자세를 인식하는 것을 목표로 한다. 여기서 사람은 한 명일 수도 있고 다수가 될 수도 있다. 또한 동작을 인식하는 데 있어 영상의 배경이 되는 장소나 사용되는 도구가 중요한 힌트가 되기도 한다. 따라서 정형화된 저차원의 feature나 rule을 사용하는 방법보다는 CNN과 같은 딥러닝을 이용하는 것이 효과적이다. 특히 본 논문에서 실험에 쓰인 UCF-101 데이터 셋의 경우, 개인이 촬영한 YouTube 비디오로 이루어져 있어 방송 컨텐츠와 같이 특정 이벤트에 한정된 구도 및 배경을 찾기 힘들다[1].

영상 내 동작 인식의 또 다른 특징은 단순 이미지 외의 정보를 활용할 수 있다는 것이다. [1]의 논문에서는 영상에서

얻을 수 있는 음성 정보를 더하여 멀티모달 시스템을 구성하였으며, [7]과 [8]에서는 영상 프레임과 Optical flow를 각각 CNN으로 학습하는 two-stream 방법으로 동작의 연속성에 따른 특징을 활용하였다. [9]의 논문에서는 CNN에 Long Short Term Memory (LSTM) 네트워크를 연결하여 동작 인식뿐 아니라 Image-to-Sentence에 응용하기도 하였다.

그러나 LSTM의 경우 네트워크의 복잡성으로 인해 한정된 데이터로는 학습이 어렵다는 단점이 있으며, [1]에서 사용한 멀티모달 시스템은 도메인이 다른 두 데이터를 fusion하여 유의미한 결과를 도출하는 데 한계가 있었다.

본 논문에서는 [1]의 시스템 성능을 향상시키기 위해서는 멀티모달 이전 단계인 이미지 검출 성능을 state-of-the-art 수준으로 끌어올려야 한다고 가정하고 다음의 학습 및 평가 방법을 제안한다.

3. 영상의 학습 및 평가 방법

프레임 이미지로부터 검출 성능을 높이기 위해 3.1과 같은 preprocessing 및 학습 데이터 추출 방법을 사용하였으며, 3.3에서 제시한 평가 방법으로 결과를 분석하였다.

실험은 GeForce GTX TitanX GPU 상에서 리눅스 기반의 Caffe[11] 및 NVIDIA DIGITS[12] 라이브러리를 사용하여 수행하였다. CNN 모델은 이미지 검출만으로 69%의 accuracy를 보인 [9]의 모델과 비교하기 위해 AlexNet을 사용하였다.

3.1. Preprocessing 및 동영상에서 학습 데이터 추출 방법

AlexNet은 227*227 크기의 이미지를 입력으로 받으므로 먼저 입력할 이미지를 resize하여 준비한다. 그러나 UCF-101 데이터 셋은 각기 다른 resolution의 영상들로 이루어져 있어, 동일한 크기로 resize 하여도 이미지에 Matte (Black bar)이 존재하게 된다. 본 논문에서는 Matte이 학습의 신뢰도를 떨어뜨리는 요인이 될 수 있다고 가정하고 그림 1과 같은 전처리 과정을 적용하였다. 이와 함께 동영상을 학습할 때 단순 키 프레임이 아닌 동작의 시간에 따른 변화 특성을 반영할 수 있도록 그림 2와 같이 여러 프레임을 혼합한 결과를 사용하였다.

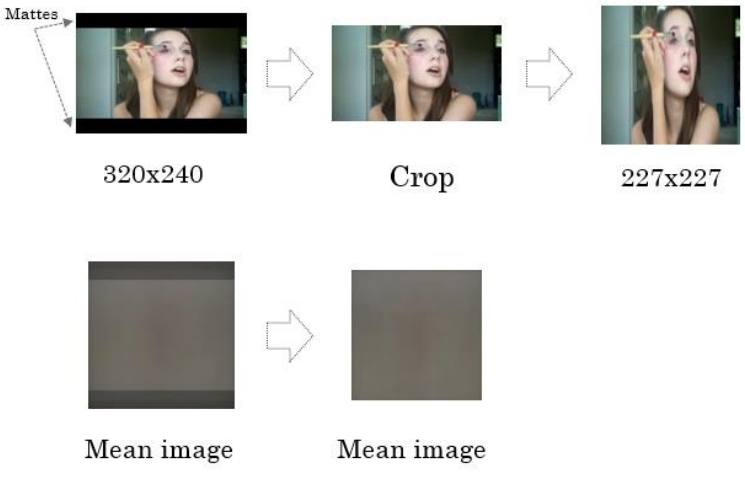


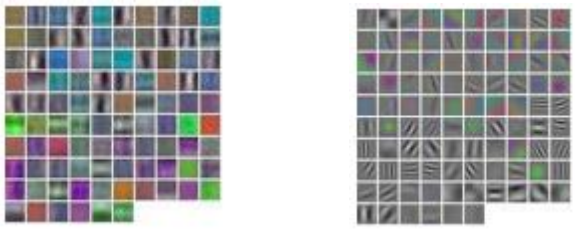
그림 1. CNN 입력을 위한 이미지 전처리

3.2. Finetuning

UCF-101 데이터 셋은 101개 클래스로 구성된 13,320개의 짧은 비디오 클립들로 Deep Neural Net을 학습하기에는 충분치 않을 수 있다. 따라서 본 실험에서는 BVLC AlexNet Caffe 모델의 weight를 초기값으로 한 Finetuning을 적용하였다. BVLC AlexNet은 1,000개 클래스로 구성된 120만장의 ImageNet LRVC-2012 데이터 셋으로 학습된 모델로 80.2%의 Top-5 accuracy를 가진다. 여기서는 Conv layer의 weight는 freeze한 상태로 마지막 Fully Connected layer 3개를 UCF-101 데이터에 대해 학습하였다.

그림 2. 동작의 시간에 따른 변화를 반영한 입력 이미지

일반적으로 잘 학습된 네트워크 필터는 깔끔한 형태의 패턴을 보이는 반면, 충분히 학습되지 않았거나 regularization 문제로 overfitting이 발생한 네트워크 필터는 노이즈가 있는 불명확한 패턴을 보인다[2]. 실제로 본 실험에서도 그림 3의 (b)와 같이 Finetuning된 네트워크에서는 high frequency gray scale feature와 low frequency color feature가 명확히 드러남을 알 수 있었다. 그에 비해 UCF-101 데이터로만 학습한 그림 3 (a)의 네트워크는 필터 간의 특징이 명확하지 않고 검출 성능도 현저히 떨어졌다.



(a) Before Finetuning (b) After Finetuning

그림 3. AlexNet Conv1 layer의 Weights 시각화

3.3. 평가 방법 및 결과 분석

학습된 모델의 테스트는 정확한 성능 평가를 위해 클래스별로 같은 수의 테스트 셋을 만들어 진행하였다. 또한 같은 모델이라도 평가 방식에 따라 계산된 결과값에 다소 차이가 있을 수 있으므로, 본 실험에서는 다음과 같은 평가 방법을 제시한다. 시스템에 테스트 영상이 입력되면 매 10번째 프레임을 추출하여 각각 CNN을 통해 영상 내 동작을 인식한다. 영상을 분류하기 위한 최종적 decision은 추출된 프레임에 대한 결과값의 평균을 이용한다. 또한 클래스 간의 검출 성능 차이가 모델의 전체 성능에 영향을 미치지 않도록

accuracy는 클래스 별로 계산하여 정규화 한다.

표 1은 위에서 제안한 학습 방법을 적용한 AlexNet으로 데이터 셋 1에 대해 테스트한 결과를 보여준다. 특히 여기서는 클래스 별 학습 데이터의 수를 일정하게 했을 경우와 모든 데이터를 학습했을 경우의 성능을 비교하였다. 그림 4에서 보듯이 UCF-101 데이터 셋은 클래스 별로 비디오 클립의 수가 달라 학습 데이터가 많은 클래스에 편향될 우려가 있다. 그러나 표 1의 테스트 결과로부터 데이터 수의 차이가 극단적이지 않은 경우, 보다 많은 데이터를 학습하는 것이 전체적인 성능에 유리하다고 추론해 볼 수 있다.

표 1. Finetuned AlexNet을 이용한 UCF-101 데이터 셋 1에 대한 실험 결과

Training data		Equal # of clips	All clips
Accuracy	Top-1	0.597	0.609
	Top-5	0.843	0.851

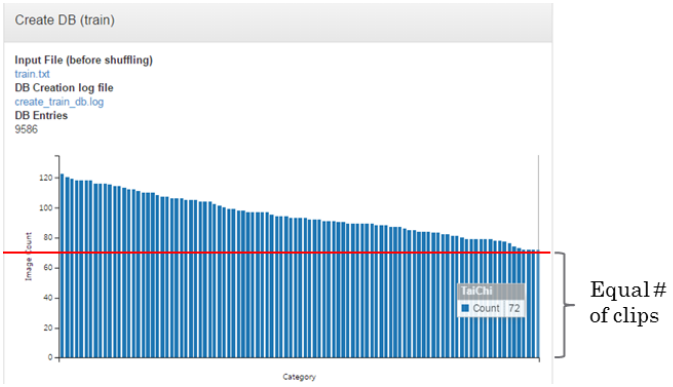


그림 4. 학습 데이터의 클래스 별 이미지 개수

표 2는 본 논문에서 제안한 학습 방법의 효과를 비교한 것이다. 본 논문의 base가 된 [1]의 시스템보다 Top-1 accuracy가 약 7.6%p 향상되었으며, 제안한 전처리 과정 없이 단순 키 프레임으로 학습한 모델보다 약 2%p 좋은 성능을 보이는 것을 확인할 수 있다. 이는 state-of-the-art라고 할 수 있는 [9]의 모델에서 이미지 정보만을 이용하여 테스트한 것과 유사한 성능이다.

표 2. UCF-101 데이터 셋에 대한 모델 별 성능 비교

Model	Base[1]	Fine-tuned	Finetune + Proposed	State-of-the-art[9]
Top-1 Accuracy	0.559	0.614	0.635	0.614 (0.69*)

* Perf. claimed from the paper[9].

4. 결론 및 향후 과제

본 논문에서는 Convolutional Neural Net으로 resolution 및 구도가 일정치 않은 영상으로부터 동작을 인식하기 위해, 먼저 이미지 정보를 이용한 검출 성능을 state-of-the-art 수준으로 만드는 것을 목표로 하였다. 결론적으로 동영상에서 학습할 이미지를 효과적으로 추출하고 Finetuning을 적용함으로써 이미지 정보만으로 63.5%의 Top-1 accuracy를 얻을 수 있었으며, 여기에 다른 영상정보를 더하여 더 높은 성능을 얻을 수 있을 것으로 기대된다.

이를 기반으로 향후에는 optical flow와 같은 영상 내 동작 표현에 적합한 motion 정보를 추가하여 다양한 비교 실험 및 분석을 할 예정이다.

참 고 문 헌

[1] Y. J. Lee et al., "A Personal Video Event Classification Method by DNN-Learning on Association between Media Modalities," *Proc. of KCC 2016*, June 2016.

[2] Stanford Vision Lab, cs231n.stanford.edu, 2016.

[3] A. Krizhevsky et al., "ImageNet Classification with Deep Convolutional Neural Networks," *Proc. of NIPS 2012*, Dec. 2012.

[4] K. Simonyan et al., "Very Deep Convolutional Networks for Large-Scale Image Recognition," *Proc. of ICLR 2015*, May 2015.

[5] C. Szegedy et al., "Going Deeper with Convolutions," *Proc. of IEEE CVPR 2015*, June 2015.

[6] K. He et al. "Deep Residual Learning for Image Recognition," *Proc. of IEEE CVPR 2016*, June 2016.

[7] K. Simonyan et al., "Two-Stream Convolutional Networks for Action Recognition in Videos," *Proc. of NIPS 2014*, Dec. 2014.

[8] H. Ye et al., "Evaluating Two-Stream CNN for Video Classification," *Proc. of ACM ICMR 2015*, June 2015.

[9] J. Donahue et al., "Long-term Recurrent Convolutional Networks for Visual Recognition and Description," *Proc. of IEEE CVPR 2015*, June 2015

[10] K. Soomro et al., "UCF101: A Dataset of 101 Human Action Classes From Videos in The Wild," *CRCV-TR-12-01*, Nov. 2012.

[11] Y. Jia et al., "Caffe: Convolutional Architecture for Fast Feature Embedding," *Proc. of the 22nd ACM International Conference on Multimedia*, pp. 675-678, 2014.

[12] NVIDIA DIGITS, https://developer.nvidia.com/digits.