비분할 비디오로부터 행동 탐지를 위한 순환 신경망 학습

송영택*, 서준배**, 김인철*
*경기대학교 컴퓨터과학과
email:dudtroc@kyonggi.ac.kr, sjb378@kyonggi.ac.kr, kic@kyonggi.ac.kr

Learning Recurrent Neural Networks for Activity Detection from Untrimmed Videos

YeongTaek Song*, Junbae Suh**, Incheol Kim*
*Dept of Computer Science, Kyonggi University

요 약

본 논문에서는 비분할 비디오로부터 이 비디오에 담긴 사람의 행동을 효과적으로 탐지해내기 위한 심층 신경망 모델을 제안한다. 일반적으로 비디오에서 사람의 행동을 탐지해내는 작업은 크게 비디오에서 행동 탐지에 효과적인 특징들을 추출해내는 과정과 이 특징들을 토대로 비디오에 담긴 행동을 탐지해내는 과정을 포함한다. 본 논문에서는 특징 추출 과정과 행동 탐지 과정에 이용할 심층 신경망 모델을 제시한다. 특히 비디오로부터 각 행동별 시간적, 공간적 패턴을 잘 표현할 수 있는 특징들을 추출해내기 위해서는 C3D 및 I-ResNet 합성곱 신경망 모델을 이용하고, 시계열 특징 벡터들로부터 행동을 자동 판별해내기 위해서는 양방향 BI-LSTM 순환 신경망 모델을 이용한다. 대용량의 공개 벤치마크 데이터 집합인 ActivityNet 비디오 데이터를 이용한 실험을 통해, 본 논문에서 제안하는 심층 신경망 모델의 성능과 효과를 확인할 수 있었다.

1. 서론

최근 스마트 폰과 디지털 카메라, 캠코더 등 저가의 고성능 비디오 장비의 대중화로 인해, 다양한 비디오 데이터의 생산과 소비가 급증하고 있다. Youtube에서는 매 분마다 약 300시간 분량의 비디오 데이터가 새로 업데이트 되고 있다고 보고된 바가 있다. 비디오 데이터가 증가함에 따라, 비디오 캡션 생성(video captioning), 비디오 기반 질의-응답(video question-answering), 사람 행동 탐지(human activity detection) 등과 같이 비디오의 내용(video content)을 자동 분석함으로써 비디오 데이터의 소비를 도와주는 편리한 기술들도 함께 주목받고 있다.

비디오 기반 사람 행동 탐지 기술은 (그림 1)의 예와 같이, 좌측 상단의 비디오 데이터 입력으로부터 우측 상단에 표시된 것과 같이 비디오에 담긴 사람의 행동이 무엇인지 추정해낼 뿐만 아니라, 하단에 표시된 것과 같이 그 행동이 포함된 비디오 영역/구간을 자동 탐지해내는 기술을 말한다. 이러한 사람 행동 탐지 기술은 환자/치매노인 생활 지원 서비스 시스템, 지능형 소셜 로봇 시스템, 영상 기반 감시 및 보안 시스템, VOD 자료 관리 시스템, 무인 자율 이동 시스템 등 실내외 환경에서 폭넓게 활용될 수 있다.

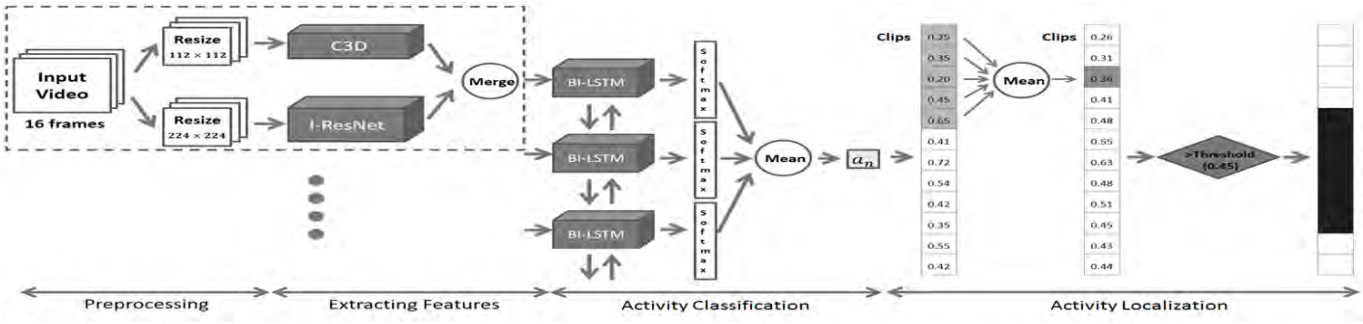
본 논문에서는 비분할 비디오(untrimmed video)로부터 사람의 행동을 효과적으로 탐지해내기 위한 심층 신경망 모델을 제안한다. 본 논문에서 제안하는 심층 신경망 모델에서는 비디오로부터 각 행동별 시간적, 공간적 패턴을 표현하는 특징(feature)들을 추출해내기 위해 C3D 및 I-ResNet 합성곱 신경망(Convolution Neural Network, CNN) 모델을 이용하고, 시계열 특징 벡터들로부터 행동을 자동 판별해내기 위해 양방향 BI-LSTM 순환 신경망(Recurrent Neural Network, RNN) 모델을 이용한다. 본 논문에서 제안하는 심층 신경망 모델의 성능과 효과를 분석하기 위해, 대용량의 공개 벤치마크 데이터 집합인 ActivityNet[1] 비디오 데이터를 이용한 실험을 수행하고 그 결과를 소개한다.

* 본 연구는 산업통상자원부의 재원으로 기술혁신사업의 지원을 받아 수행한 연구과제 (No. 10060086, 개인 서비스용 로봇을 위한 지능-지식 집약, 개방, 진화형 로봇지능 소프트웨어 프레임워크 기술 개발)입니다.

(그림 2) 비디오 기반 사람 행동 탐지의 예

2. 관련연구

일반적으로 비디오에서 사람의 행동을 탐지해내는 작업은 크게 비디오에서 행동 탐지에 효과적인 특징들을 추출해내는 과정(feature extraction)과 이 특징들을 토대로 비디오에 담긴 행동을 탐지해내는 과정(activity detection)을 포함한다. 그리고 이러한 행동 탐지 과정은 다시 비디오에 담긴 행동 유형을 판별하는 행동 분류 과정(activity classification)과 비디오 내에서 해당 행동이 등장하는 시



(그림 3) 비디오 행동 탐지를 위한 전체 과정

간적 영역을 판별하는 행동 영역 탐지 과정(activity localization)으로 나뉜다. 먼저, Gurkirt의 연구[2]에서는 전체 비디오를 대상으로 행동 탐지에 필요한 단일 점수 벡터(single score vector)를 추출해낸다. 단일 점수 벡터는 전체 비디오를 대상으로 먼저 ImageNetShuffle, MBH (Motion Boundary Histogram), C3D(3D Convolutional Neural Network)[3] 등의 특징들을 추출하고, 각각의 특징 별로 사전에 일대다(one versus rest) 방식으로 학습시킨 SVM(Support Vector Machine) 모델을 적용하여 서로 다른 3가지 SVM 점수를 얻어내고 이들을 하나의 벡터로 결합한 것이다. 행동 탐지를 위해서는 단일 점수 벡터를 토대로 다시 일대다 SVM 행동 분류 모델을 학습시킨 뒤, 이를 이용하여 전체 비디오에 담긴 행동을 분류해낸다. 행동 영역 탐지(activity localization)를 위해서는 C3D 특징으로 이진 랜덤 포레스트(binary random forest)를 학습시키고 이를 이용하여 개별 프레임 단위로 행동 영역 여부를 판별한다. 이 연구에서 제시한 방법은 행동 분류를 위해 추출한 C3D 특징을 행동 영역 탐지에도 효과적으로 재사용한다는 장점은 가지고 있으나, 전체 비디오의 길이가 길어질수록 다수의 행동이 포함될 수 있기 때문에 높은 분류 정확도를 얻기 어렵다는 한계점을 가진다. 한편, Montes의 연구[4]에서는 전체 비디오를 일정한 구간 별로 나누고 각 구간을 대상으로 C3D 특징을 추출해낸다. 행동 분류를 위해서는 추출된 C3D 특징을 토대로 순환 신경망의 하나인 LSTM(Long Short Term Memory)을 학습시키고, 이를 이용하여 구간별 행동을 분류한다. 행동 분류를 위해 LSTM 순환 신경망을 이용하는 이 방법은 연속된 구간 단위의 시계열 패턴을 학습함으로써 행동 분류의 정확도는 어느 정도 향상시킬 수 있었으나, 행동 분류를 위해 C3D 특징이외에 보다 다양한 특징들을 이용하지 못했다는 한계점이 있다. 또한, Limin의 연구[5]에서는 전체 비디오를 일정한 구간으로 분할하고 각 구간 별로 TSN(Temporal Segment Networks)를 이용하여 광 흐름(optical flow) 평균 특징과 RGB 평균 특징을 추출해낸다. 그리고 완전 연결 계층(fully connected layer)들로 구성된 분류 신경망 모델을 이용해 각 특징별 분류 점수를 구해내고 이들을 혼합함으로써, 비디오 전체에 담긴 사람의 행동을 분류해낸다. 이 방법은 행동 분류를 위해 광 흐름 특징과 RGB 특징 등 서로 다른 특징들을 보완적으로 이용한다는 장점은 있으나, 채용하고 있는 분류 모델의 특성 상 각 구간의 행동 분류에 인접 구간들의 맥락 정보(context)를 효과적으로 반영할 수 없다는 단점이 있다.

3. 순환 신경망 기반의 비디오 행동 탐지

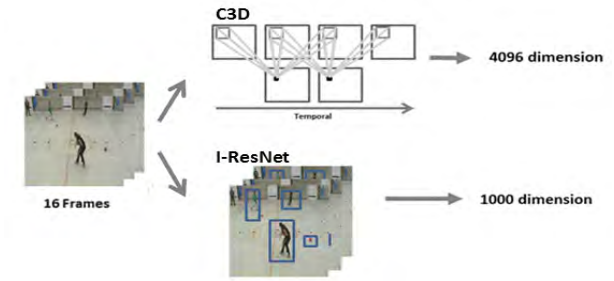
본 논문에서 제안하는 순환 신경망 기반의 비디오 행동 탐지의 전체 과정은 (그림 2)와 같다. 비디오 행동 탐지 과정은 크게 전처리(preprocessing), 특징 추출(extracting features), 행동 분류(activity classification), 행동 영역 탐지(activity localization) 과정으로 구성된다. 전처리 과정에서는 전체 비디오를 16 프레임(16 frame) 단위의 각 구간들로 나누고, 특징 추출을 위한 두 합성곱 신경망(C3D, I-ResNet)의 입력 형식에 맞추어 각 구간 비디오의 크기를 조절(resize)한다. 특징 추출 과정에서는 각 구간 비디오에서로 다른 합성곱 신경망 모델들인 C3D와 I-ResNet을 적용하여 상호 보완적인 특징들을 추출해내고, 이들을 하나의 특징 벡터로 병합(merge)한다. 행동 분류 과정에서는 순환 신경망의 하나인 양방향 BI-LSTM 모델을 이용하여 16 프레임 단위의 각 구간 비디오에서 추출한 특징 벡터들의 시퀀스로부터 시계열 패턴을 학습하고, 이를 토대로 각 구간 비디오의 행동별 분류 점수를 계산해낸다. 그리고 이러한 각 구간별 분류 점수를 토대로 최종 비디오 행동(a_n)을 판별해낸다. 마지막으로 행동 영역 탐지 과정에서는 각 구간별로 행동(a_n)의 분류 점수가 미리 정해진 임계값(threshold)보다 높은 경우에 이 구간을 행동(a_n)이 발생한 시간 영역으로 판별한다.

3.1 전 처리

본 연구에서는 전체 비디오를 일정한 크기의 구간 비디오들로 나누고, 특징 추출을 위한 구간 비디오의 크기 조절을 위해 전처리 과정이 필요하다. 본 연구에서는 전체 비디오를 16 프레임 단위의 구간 비디오들로 나눈다. 그리고 분할된 구간 비디오들을 특징 추출용 합성곱 신경망인 C3D에 입력하기 위해서는 각 프레임을 112 X 112 크기로 조절하며, I-ResNet에 입력하기 위해서는 224 X 224 크기로 조절한다.

3.2 특징 추출

전처리가 완료되면, 크기가 조절된 두 구간 비디오를 각각 합성곱 신경망 모델인 C3D와 I-ResNet에 입력하고 그 결과를 병합함으로써 구간별 특징 벡터를 생성한다. 상세한 특징 추출 과정은 (그림 3)과 같다. 먼저, (그림 3)의 합성곱 신경망 C3D는 112 X 112 크기의 프레임들로 구성된 구간 비디오로부터 4096 차원의 특징 벡터를 출력한다. 많은 선행 연구들을 통해 2차원 합성곱 신경망 모델인 2D-CNN(Convolutional Neural Network)은 각 프레임 단위의 영상(image)으로부터 해당 영상에 담긴 물체의 분류와 탐지를 위한 특징들을 자동 학습하는데 매우 효과적이라고 알려져 있는데 반해, 이를 확장한 3차원 합성곱 신경망 모델인 C3D는 프레임들의 시퀀스인 비디오와 같은 3차원 데이터로부터 해당 비디오에 담긴 행동을 분류하고 탐지하는데 매우 효과적이라고 알려져 있다. C3D 특징은 각 프레임 영상 내 공간적인 정보뿐만 아니라 프레임들 간의 연속된 시간적인 정보도 포함할 수 있기 때문에, 하모니카 연주와 같이 연속된 여러 프레임들에 걸쳐 이루어지는 하나의 행동을 포착해내는데 특별히 강점이 있어 채택하였다.

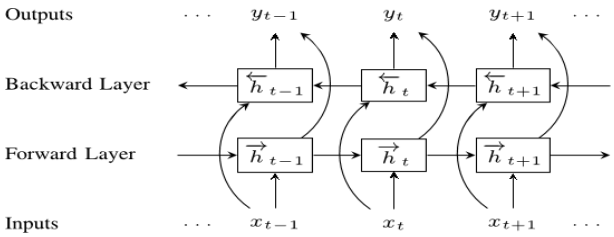


(그림 4) 합성곱 신경망 모델 기반의 특징 추출

한편, 구기 스포츠, 악기 연주, 집안 청소 등과 같은 많은 일상 행동들은 대부분 특화된 대상 물체(object)나 도구(tool)들과 함께 등장하는 경우가 많다. 따라서 비디오에서 이러한 대상 물체나 도구들을 탐지하면 연관된 행동들을 찾아내는데 큰 도움을 줄 수 있다. 본 연구에서는 비디오에 담긴 행동을 탐지하기 위해 비디오에 등장하는 물체들을 직접 탐지하고 이용하는 대신, 또 다른 합성곱 신경망 모델인 I-ResNet을 이용해 물체 분류와 탐지에 효과적인 특징들을 추가로 추출해내고 이들을 C3D 특징들과 상호 보완적으로 이용하도록 설계하였다. 본 연구에서 이용하는 I-ResNet 합성곱 신경망 모델은 224 X 224 크기의 프레임들로 구성된 구간 비디오로부터 1000 차원의 특징 벡터를 출력한다. 특징 추출에 이용되는 I-ResNet 모델은 대표적인 합성곱 신경망의 하나인 ResNet을 대용량의 영상 데이터 집합인 ImageNet으로 미리 학습시켜 얻는다. ImageNet에는 일상생활에 등장하는 수많은 물체 영상 데이터들이 포함되어 있다.

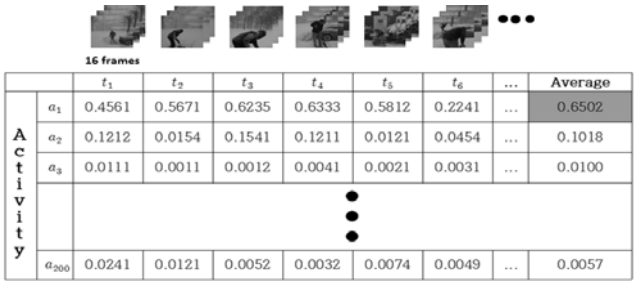
3.3 행동 분류

본 연구에서는 비디오의 각 구간에서 추출한 특징 벡터들의 시퀀스를 토대로 비디오에 포함된 사람의 행동을 분류해내기 위해, 시계열 패턴 학습에 유리한 순환 신경망 모델을 이용한다. 특히 본 연구에서는 선행 연구들에서 시도되었던 단방향(uni-directional) LSTM 순환 신경망 대신 새로운 양방향(bi-directional) LSTM, 즉 BI-LSTM 순환 신경망 모델을 채용하였다.



(그림 5) 양방향 LSTM 순환 신경망 모델

BI-LSTM은 (그림 4)와 같이 기존의 단방향 LSTM에 역방향의 은닉 층을 추가함으로써, 시계열 데이터에서 지나간 과거 입력 데이터와 상태 정보뿐만 아니라 미래의 것들까지 모두 참고할 수 있어, 단방향 LSTM에 비해 보다 풍부한 시간적 맥락 정보(temporal context)를 출력 결정에 이용할 수 있다는 장점이 있다. 비디오의 각 구간 특징 벡터 입력에 대해 BI-LSTM이 생성하는 출력은 다시 완전 연결 Softmax 층을 거쳐 해당 구간의 행동별 분류 점수로 변환된다. 그리고 각 구간에서 구한 행동별 분류 점수들의 평균을 비교해봄으로써 전체 비디오에 담긴 최종 행동이 결정된다. (그림 5)는 16 프레임 단위의 각 비디오 구간($t_1 ... t_n$)에 대해, BI-LSTM과 Softmax 층이 생성하는 행동별($a_1 ... a_{200}$) 분류 점수와 평균의 예를 보여준다. (그림 5)의 예에서는 각 구간의 행동별 분류 점수의 평균치(average)가 0.6502로 가장 높은 a_1를 이 전체 비디오에 담긴 주된 사람의 행동으로 판단한다.

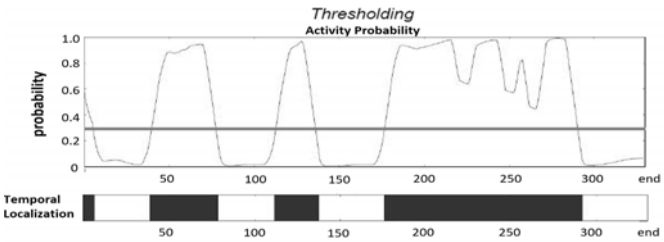


(그림 6) 각 구간(t_i)의 행동별(a_j) 분류 점수

행동 분류를 위해 본 연구에서 채택한 BI-LSTM은 기존의 단방향 LSTM에 비해 보다 풍부한 시간적 맥락정보를 이용할 수 있으므로 더 높은 행동 분류 성능 향상을 기대할 수 있는 반면, 단방향 LSTM에 비해 더 많은 내부 파라미터(parameter)들을 포함하고 있기 때문에 좀 더 많은 학습 데이터와 학습 시간을 요구할 것으로 예상한다.

3.4 행동 영역 탐지

행동 영역 탐지(activity localization)는 분류 과정을 통해 알아낸 비디오에 담긴 사람의 특정 행동(a^*)이 실제로 등장하는 비디오 구간들을 찾아내는 작업이다. 본 연구에서는 행동 분류 과정동안 비디오의 각 구간에서 계산된 행동 a^*의 분류 점수들을 먼저 평활화(smoothing)한 다음, 평활화된 행동 a^*의 분류 점수가 미리 정의해놓은 임계값(threshold)보다 더 높은 비디오 구간들은 모두 행동 a^*가 발생한 영역으로 판단한다.



(그림 7) 임계값 기반의 행동 영역 탐지

(그림 6)의 상단은 각 구간에서 계산한 행동 a^*의 평활화된 분류 점수(activity score/probability)를 그래프 형태로 보여주고 있다. 그리고 가로선으로 표시된 임계값보다 분류 점수가 더 높은 구간들은 (그림 6)의 하단에 음영으로 표시된 영역들과 같이 모두 비디오에서 행동 a^*가 발생한 영역으로 판별한다.

4. 구현 및 실험

4.1 데이터 집합

본 연구에서 사용한 비디오 데이터 집합은 ActivityNet Challenge 2016[4]에서 제공하는 비디오 행동 탐지용 벤치마크 데이터 집합인 ActivityNet200이다. ActivityNet200은 계층화된 200개의 서로 다른 일상 행동 클래스들을 정의하고 있으며, 각 행동 클래스에 속한 약 10,024개의 학습용 비디오, 4,926개의 검증용 비디오, 5,044개의 테스트용 비디오를 포함하고 있다. 각 비디오마다 평균 1.65개의 행동 인스턴스를 포함하고 있으며, 총 용량은 440GB이다.

4.2 구현 및 학습

실험을 위해 Ubuntu 14.04 UTS 환경에서 Python 딥러닝 라이브러리인 Keras를 기반으로 본 논문에서 제안하는 심층 신경망 모델을 구현하였다. 모델 학습과 실험은 4.0Ghz 4Core, 8Thread CPU와 Geforce GTX TITAN X GPU카드가 설치된 하드웨어 환경에서 수행하였다. 모델 최적화 알고리즘은 순환 신경망 학습에 적합한 RMSprop을, 일괄 처리량(batch size)은 256, 반복 횟수(epoch)는 20, 학습률(learning rate)은 10^{-5}으로 각각 설정한 후, 모델 학습을 수행하였다.

4.3 실험 및 평가

첫 번째 실험에서는 C3D 단일 특징 모델과 본 논문에서 제안한 C3D+I-ResNet 합성 특징 모델에 따른 학습 및 행동 분류 성능을 비교 분석해보았다.

<표 1> 특징 모델에 따른 학습 및 분류 성능 비교

classifier / measure	LSTM		BI-LSTM	
	C3D	C3D + I-ResNet	C3D	C3D + I-ResNet
Time	108m	130m	38m	47m
Epoch	100	100	20	20
Clip_acc	0.4878	0.5248	0.5140	0.5426
Vid_acc	0.5094	0.5349	0.5118	0.5522

Learning rate = 1e-5

<표 1>은 C3D 단일 특징 모델과 C3D+I-ResNet 합성 특징 모델의 성능을 비교한 첫 번째 실험 결과를 나타내며, 표에서 서로 다른 두 분류기(classifier)인 LSTM과 BI-LSTM을 사용했을 때 각각 소요된 학습 시간(Time), 반복 횟수(Epoch), 구간별 행동 분류 정확도(Clip Accuracy), 비디오의 행동 분류 정확도(Video Accuracy) 등을 보여주고 있다. <표 1>의 결과에서, 두 분류기 모두 본 논문에서 제안한 C3D+I-ResNet 합성 특징을 이용한 경우가 C3D 단일 특징만을 이용한 때보다 비록 학습 시간은 소폭 증가하였으나, 구간별 행동 분류 정확도와 비디오 행동 분류 정확도 면에서 모두 뚜렷한 성능 개선이 있었음을 확인할 수 있다.

두 번째 실험에서는 특징 추출을 위해 C3D+I-ResNet 합성 모델을 이용하는 경우, 본 논문에서 제안하는 양방향 BI-LSTM 모델을 포함해 서로 다른 분류 모델들 간의 학습 및 행동 분류 성능을 비교 분석해보았다.

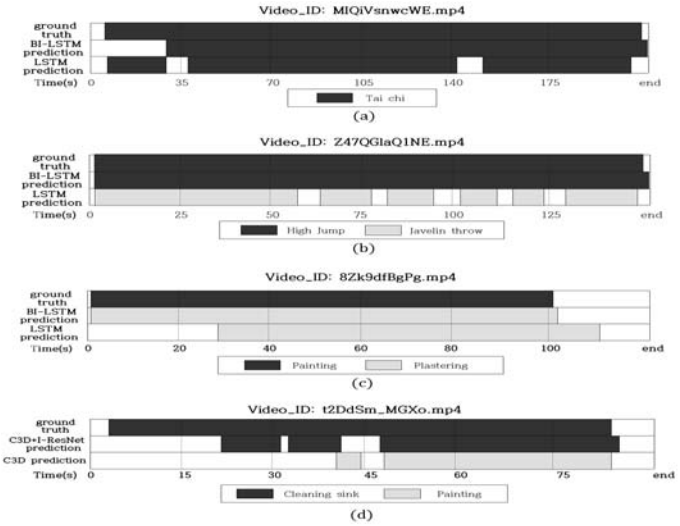
<표 2> 분류 모델에 따른 학습 및 분류 성능 비교

classifier / measure	Fc-softmax	RNN	LSTM	BI-LSTM
Time	108m	130m	130	47m
Epoch	100	100	100	20
Clip_acc	0.4857	0.4813	0.5248	0.5426
Vid_acc	0.5127	0.5013	0.5349	0.5522

Learning rate = 1e-5

<표 2>는 완전 연결(FC)-Softmax 모델, 기본 RNN 모델, 단방향 LSTM 모델, 양방향 BI-LSTM 모델 등 총 4가지 분류 모델의 성능을 비교한 두 번째 실험 결과를 나타낸다. <표 2>의 결과에서, 양방향 BI-LSTM 모델의 분류 성능이 다른 분류 모델들의 경우에 비해 가장 높게 나타났을 뿐만 아니라, 모델 학습에 소요된 학습 시간도 가장 짧게 측정되었다. 이러한 실험 결과는 양방향 BI-LSTM 모델의 높은 시계열 패턴 학습 능력과 시간 맥락성이 반영된 결과로 해석된다.

마지막 세 번째 실험에서는 분류 모델에 따른 행동 영역 탐지 성능을 비교 분석해보았다. 이 실험에서는 (a)~(c)는 비디오 특징 추출을 위해 C3D+I-ResNet 합성 모델을 사용하였고, 행동 분류 모델로는 단방향 LSTM과 양방향 BI-LSTM 모델을 비교하였다. 실험 결과의 일부를 나타내는 (그림 7)의 (a)~(c)는 각각 서로 다른 3 개의 비디오에 대해, 실제 행동 영역(Ground Truth)을 기준으로 양방향 BI-LSTM 모델의 탐지 영역과 단방향 LSTM 모델의 탐지 영역들을 비교하여 보여주고 있다. (그림 7)의 (a)~(c)의 결과에서 보듯이, 본 논문에서 제안하는 양방향 LSTM 분류 모델이 행동 영역 탐지 성능 면에서도 단방향 LSTM 모델에 비해 더 우수한 것을 확인할 수 있었다. 한편, (그림 7)의 (d)의 경우에는 C3D 특징만을 사용한 BI-LSTM 분류 모델의 탐지 영역과 C3D + I-ResNet 합성 특징을 사용한 BI-LSTM 분류 모델의 탐지 영역을 비교하여 보여주고 있다. (그림 7)의 (d)의 결과에서 볼 수 있듯이, C3D+I-ResNet의 특징을 사용하는 모델이 행동 영역 탐지 성능 면에서 더 우수한 것을 확인할 수 있다.



(그림 7) 분류 모델에 따른 행동 영역 탐지 성능 비교

5. 결론

본 논문에서는 비분할 비디오로부터 이 비디오에 담긴 사람의 행동을 효과적으로 탐지해내기 위한 심층 신경망 모델을 제시하였다. 특히 비디오로부터 특징 추출을 위해서는 합성곱 신경망 모델인 C3D+I-ResNet 합성 특징 모델을, 시계열 특징 벡터들로부터 행동을 자동 판별해내기 위해서는 순환 신경망 모델의 하나인 양방향 BI-LSTM 분류 모델을 제안하였다. 대용량의 공개 벤치마크 데이터 집합인 ActivityNet 비디오 데이터를 이용한 실험을 통해, 본 논문에서 제안하는 심층 신경망 모델의 성능과 효과를 확인할 수 있었다.

참고문헌

[1] Fabian Caba Heilbron, et al, "ActivityNet: A Large-Scale Video Benchmark for Human Activity Understanding" Proc. of CVPR-15, pp.961-970, 2015.

[2] G. Singh and F. Cuzzolin, "Untrimmed Video Classification for Activity Detection: submission to ActivityNet Challenge." arXiv preprint arXiv:1607.01979, 2016.

[3] D. Tran and L. Bourdev, et al, "Learning Spatiotemporal Features with 3D Convolutional Networks" Proc. of ICCV-15, pp.4489-4497, 2015.

[4] A. Montes, A. Salvador and S. Pascual, et al, "Temporal Activity Detection in Untrimmed Videos with Recurrent Neural Networks" arXiv preprint arXiv:1608.08128, 2016.

[5] L. Wang, Y. Xiong and Z. Wang, et al, "Temporal Segment Networks: Towards Good Practices for Deep Action Recognition." Proc. of ECCV-16, pp.20-36, 2016.