역재생 비디오를 이용한 완전 지도 시간적 행동 검출

권희원, 조혜정, 윤수용, 정찬호

국립한밭대학교

20191515@edu.hanbat.ac.kr, 20201566@edu.hanbat.ac.kr, 20181435@edu.hanbat.ac.kr, peterjung@hanbat.ac.kr

Fullly Supervised Temporal Action Localization Using Reverse Playback Videos

Huiwon Gwon, Hyejeong Cho, Suyoung Yun, Chanho Jung

Hanbat National University

요 약

최근 시간적 행동 검출 연구가 활발히 진행되고 있다. 시간적 행동 검출 연구에서 오프라인 행동 검출은 과거, 현재 및 미래의 정보를 활용할 수 있다. 이에 의해 오프라인 행동 검출은 순방향 정보와 역방향 정보 모두 사용 가능하지만, 순방향 정보만을 대부분 활용한다. 본 논문은 THUMOS-14 데이터 셋[1]을 가공하여 생성한 역재생 비디오를 통해 완전지도 시간적 행동 검출 연구를 수행한다. TALLFomer[2] 모델에 대해 순재생 비디오와 역재생 비디오를 각각 학습시키고, 이에 대한 두 결과를 정성적, 정량적으로 비교하였다. 정량적 결과 비교를 위해 mAP(mean Average Precision)를 측정하였고, 두 경우 유사한 성능을 보였다. 정성적 결과는 각각의 모델이 예측한 시간적 행동 구간을 출력하였고, 비교 결과 차이가 있음을 확인하였다.

Ⅰ. 서론

미디어 플랫폼이 발전하며 비정형 비디오에 대한 접근 및 수집이 간편해져 시간적 행동 검출 연구가 주목받기 시작했다. 시간적 행동 검출은 오프라인 행동 검출과 온라인 행동 검출로 분류된다. 오프라인 행동 검출은 비정형 비디오에서 행동의 시작과 끝을 찾고 행동 분류를 수행한다. 오프라인 행동 검출은 비디오 내에서 현재와 과거의 정보만을 활용할 수 있는 온라인 행동 검출과 달리 미래의 정보를 함께 활용한다. 이를 통해 오프라인 행동 검출 연구에서는 과거에서 현재, 현재에서 미래로 진행되는 순방향 정보를 잘 활용해 왔다. 이에 반해 미래에서 현재 및 과거로 진행되는 역방향 정보를 오프라인 행동 검출에 순방향 정보와 함께 사용하는 연구는 활발히 진행되지 않고 있다. 이는 온라인 행동 검출에 비해 많은 더욱 많은 정보를 이용할 수 있는 오프라인 행동 검출의 이점을 살리지 못한 것으로 판단된다. 순방향 및 역방향 정보 모두 사용하는 Bi-directional LSTM을 통해 오프라인 행동 검출 연구[3, 4]가 진행되었지만 이후 지속적인 연구가 이루어지지 않았다.

오프라인 행동 검출은 지도방식에 따라 완전지도 시간적 행동 검출, 약지도 시간적 행동 검출, 포인트 주석 기반 약지도 시간적 행동 검출로 나누어진다. 그중 완전지도 시간적 행동 검출은 비정형 비디오 내 행동의 시작과 끝, 행동의 종류를 어노테이션으로 제공한다.

역재생 비디오를 사용해 TALLFomer[2] 모델 학습을 진행하여 이를 통해 비디오가 담고 있는 역방향 정보를 쉽게 학습할 수 있다. 이에 본 논문에서는 역재생 비디오가 담고 있는 역방향 정보의 가치를 확인하고 양방향 정보를 활용한 오프라인 행동 검출에 대한 연구로 이어나가고자 역재생 비디오를 사용한 완전지도 시간적 행동 검출 연구를 진행한다. 연구를 위해 THUMOS-14 데이터 셋[1]의 학습 데이터와 평가 데이터 각각에 대한 역재생 비디오를 추가적으로 생성한다. TALLFomer[2] 모델에 대한 비디오 재생 방향에 따른 입력을 달리하며 학습 및 평가를 진행하고, 이 결과를 정량적, 정성적으로 비교한다.

Ⅱ. 본론

데이터 셋. THUMOS-14 데이터 셋[1]은 200개의 검증 데이터와 212개의 평가 데이터를 가진다. 기존의 완전지도 시간적 행동 검출 연구의 기조를 따라 200개의 검증 데이터를 학습 데이터로 사용하여 연구를 진행한다.

역재생 비디오. 순재생 비디오로 구성된 학습 데이터, 평가 데이터에 대한 역재생 비디오를 추가로 생성한다. 완전 지도 시간적 행동 검출 연구를 진행하기 위해서 비디오 내 행동의 정확한 시작 시점과 종료 시점을 모델에 입력해 주어야 한다. 역재생 비디오 내 행동의 시작 시점과 종료 시점의 정의는 아래와 같다.

$$\begin{aligned} Rev\_st &= |\, Fwd\_et - Last\_Time \,| \\ Rev\_et &= |\, Fwd\_st - Last\_Time \,| \end{aligned} \quad (1)$$

식 (1)에서 Rev_st, Rev_et와 Fwd_st, Fwd_et는 각각 역재생 비디오와 순재생 비디오 내 행동의 시작 시점과 종료 시점을 나타낸다. Last_time은 순재생 비디오 1개의 총 재생 시간이다.

실험 구성. 비디오 재생 방향에 따라 학습 및 평가 데이터를 달리했을 때 TALLFomer[2] 모델에 대한 학습 결과를 비교하기 위해 표1 과 같이 실험을 구성한다. 표 1에서 보이는 바와 같이 (가), (나) 실험은 학습 데이터와 평가 데이터가 동일한 재생 방향을 가진 비디오이며, 각각 역재생 비디오와 순재생 비디오를 사용한다. (다), (라) 실험은 학습 데이터와 평가 데이터를 서로 다른 재생 방향을 가진 비디오로 구성한다. (다)는 역재생 비디오, (라)는 순재생 비디오를 학습 데이터로 사용한다. 실험 결과를 출력하고 (가)와 (나) 및 (다)와 (라)를 비교한다.

표 1. 모델 입력 데이터에 따른 실험 구성

모델	방법	학습 데이터	평가 데이터
TALLFormer[2]	(가)	역재생 비디오	역재생 비디오
	(나)	순재생 비디오	순재생 비디오
	(다)	역재생 비디오	순재생 비디오
	(라)	순재생 비디오	역재생 비디오

표 2. 입력 데이터에 따른 TALLFormer[2] mAP 성능 비교

모델	방법	mAP@IoU(%)			AVG
		0.3	0.5	0.7	(0.3:0.7)
TALLFormer[2]	(가)	76.6	**63.3**	**34.7**	59.3
	(나)	**78.2**	63.2	34.4	**60.0**
	(다)	57.0	**36.6**	**12.8**	35.5
	(라)	**58.4**	36.0	11.6	35.5

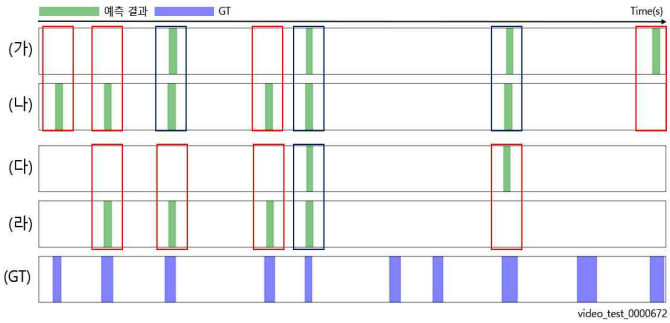

그림 1. TALLFormer[2] 672번 평가 비디오에 대해 예측한 시간적 행동 구간 : 빨간 박스 - (가), (나) 또는 (다), (라) 결과가 다른 경우, 파란 박스 : (가), (나) 또는 (다), (라) 결과가 같은 경우

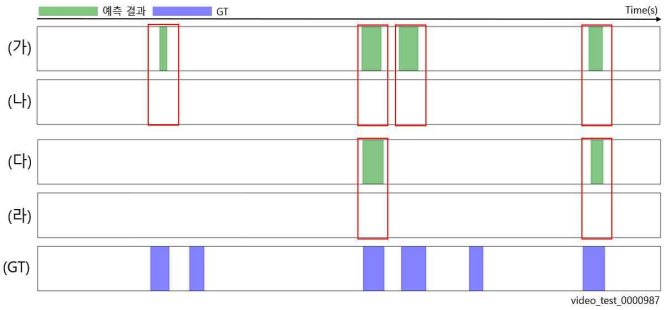

그림 2 TALLFormer[2]가 987번 평가 비디오에 대해 예측한 시간적 행동 구간: 빨간 박스 - (가), (나) 또는 (다), (라) 결과가 다른 경우, 파란 박스 : (가), (나) 또는 (다), (라) 결과가 같은 경우

정량적 결과. 표 2는 표 1에 정리된 것과 같이 TALLFormer[2]에 대한 실험을 진행하고 얻은 성능 비교를 보여준다. tIoU(temporal Intersection over Union)의 임계값을 변화시켜 가며 mAP(mean Average Precision)를 측정한다. AVG는 tIoU 임계값 0.3:0.1:0.7에서 평균 mAP를 나타낸다.

성능 비교. 표2에서 보는 바와 같이 (가, 나)학습 데이터 및 평가 데이터의 재생 방향이 같은 경우 사용한 비디오의 재생 방향에 상관없이 평균 mAP가 유사함을 보여주었다. 또한 (다, 라)학습 데이터 및 평가 데이터의 재생 방향이 다를 경우 학습 비디오의 재생 방향에 상관없이 평균 mAP의 성능이 동일함을 알 수 있었다.

정성적 결과. 그림 1, 그림 2는 표 1의 실험을 통해 TALLFormer[2]가 예측한 시간적 행동 구간을 서로 다른 비디오에 대해 시각화한 결과이다. 각각의 시간적 행동 구간은 행동 신뢰 점수가 0.7 이상을 가지며 행동 분류에도 성공한 경우다. (가, 라)역재생 비디오를 평가 데이터로 사용하여 시간적 행동 구간을 출력한 경우 결과값을 반전하여 (나, 다) 순재생 비디오에 대해 평가한 결과와 비교를 진행하였다.

그림 1, 그림 2 공통점. 학습 데이터와 평가 데이터의 비디오 재생 방향이 같은 경우를 비교하였을 때 (가)역재생 비디오를 이용한 경우에 시간적 행동 구간 예측에 성공하였지만 (나)순재생 비디오를 이용한 경우에서 실패한 구간이 존재한다. 또한 학습 데이터와 평가 데이터의 재생 방향이 다른 경우를 비교하였을 때 학습 데이터를 (다)역재생 비디오로 구성한 경우에서 시간적 행동 구간 예측에 성공하였지만 (라)순재생 비디오로 구성한 경우에서 실패한 구간이 존재한다.

그림 1, 그림 2 차이점. 그림 1는 662번 평가 비디오에 대해 모델이 예측한 시간적 행동 구간을 나타내며 (나, 라)순재생 비디오를 학습 데이터로 사용한 경우 더 많은 모델 예측 결과를 보여준다. 그림 2는 986번 평가 비디오에 대해 모델이 예측한 시간적 행동 구간을 나타내며 (가, 다)역재생 비디오를 학습 데이터로 이용한 경우에만 모델이 예측에 성공하였다.

분석. 성능 비교를 통해 역재생 비디오와 순재생 비디오를 각각 모델에 학습시켰을 때 유사한 평균 mAP 성능을 가지는 것을 확인하였다. 이를 통해 역재생 비디오가 완전지도 시간적 행동 검출 모델에 제공하는 역방향 정보의 유효성을 확인할 수 있다. 또한 그림 1, 그림 2의 비교를 통해 역재생 비디오의 역방향 정보를 학습시킨 모델은 순방향 정보를 학습시킨 모델과 다른 방식으로 시간적 행동 구간을 예측함을 보여준다.

Ⅲ. 결론

본 논문에서는 역재생 비디오를 사용하여 완전지도 시간적 행동 검출 연구를 수행하였다. TALLFormer[2]에 역재생 비디오를 학습시킨 결과를 순재생 비디오 학습 결과와 정성적, 정량적으로 비교하였다. 정량적 결과 비교를 통해 역방향 정보가 순방향 정보만큼 유의미함을 알았다. 또한 정성적 결과 비교를 통해 순방향 정보와 차별성이 있음을 확인하였다. 본 논문을 통해 양방향 정보를 활용한 오프라인 행동 검출 연구로 이어나갈 수 있을 것이라 판단된다.

참 고 문 헌

[1] Y.-G. Jiang, J. Liu, A. R. Zamir, G. Toderici, I. Laptev, M. Shah, and R. Sukthankar.“ Thumos challenge: Action recognition with a large number of classes,” 2016.

[2] Cheng, Feng and Bertasius, Gedas.” TALLFormer: Temporal Action Localization with Long-memory Transformer,” ECCV, pages 503 - 521, 2022.

[3] Tianwei Lin, Xu Zhao*, Zhaoxuan Fan.” TEMPORAL ACTION LOCALIZATION WITH TWO-STREAM SEGMENT-BASED RNN,” IEEE, pages 2381-8549, 2017.

[4] Bharat Singh, Tim K. Marks, Michael Jones, Oncel Tuzel, Ming Shao.” A Multi-Stream Bi-Directional Recurrent Neural Network for Fine-Grained Action Detection,” CVPR, pages 1961-1970, 2016.