논문 20-45-06-05 The Journal of Korean Institute of Communications and Information Sciences '20-06 Vol.45 No.06
https://doi.org/10.7840/kics.2020.45.6.975

시간적 행동 구간 생성을 위한 LSTM 기반 단편 관련성 학습

은 현 준◆, 문 진 영*, 박 종 열*,
정 찬 호°, 김 창 익**

Learning Snippet Relatedness Based on LSTM for Temporal Action Proposal Generation

Hyunjun Eun◆, Jinyoung Moon*,
Jongyoul Park*, Chanho Jung°,
Changick Kim**

요 약

최근 많은 미디어 플랫폼의 발달로 비정형 비디오들의 수집과 접근이 용이해졌다. 이에 따라 비디오 이해를 위해 비정형 비디오에서 행동의 시작과 끝을 찾는 시간적 행동 검출 연구가 최근 활발히 이루어지고 있다. 시간적 행동 구간 생성 방법은 Temporal Convolutional Network를 이용하여 행동 구간을 정의한다. 이와는 다르게, 본 논문에서는 행동을 시간적 발생 순서에 따라 모델링 하기 위하여 LSTM을 이용한 방법을 제안한다. 제안하는 방법은 LSTM을 이용하여 단편 관련성 (Snippet Relatedness)를 평가하고 이를 통해 행동 구간을 정의한다. 단편 관련성은 단편들이 서로 동일한 행동 구간에 포함되는지를 나타내는 지표이다. 제안하는 방법은 THUMOS-14 데이터 셋에 대한 실험에서 50개의 행동 구간 수 추출 시 41.34% 평균 리콜 성능을 얻어 BSN, MGG 보다 3.88%, 1.41% 우수한 성능을, SRG보다는 0.85% 떨어진 성능을 보였다.

Key Words : temporal action proposal generation, snippet relatedness, LSTM

ABSTRACT

Recent temporal action proposal generation approaches are based on temporal convolutional networks. In this paper, different from this, we propose to use LSTM for sequential modeling on actions. The propose method based on LSTM evaluates snippet relatedness to define temporal action intervals. Snippet relatedness indicates which snippets are included in the same action instance. By conducting experiments on the THUMOS-14 dataset, we demonstrate the superiority of the proposed method. We also analyze our method in diverse aspects.

Ⅰ. 서 론

최근 많은 촬영 장비 및 미디어 플랫폼의 발달로 사람들은 쉽고 빠르게 많은 비디오를 접할 수 있게 되었다. 이에 따라 실생활에서 촬영되는 비정형 비디오 관련 분석 연구[1]도 활발히 이루어지고 있으며, Temporal Action Proposal Generation도 그중 하나라고 볼 수 있다. Temporal Action Proposal Generation은 비정형 비디오에서 행동 구간을 찾는 작업으로 행동의 시작 시간, 끝 시간, 행동 신뢰 점수를 평가한다. 비정형 비디오에서 찾아진 행동 구간은 비디오 이해 테스크 중 하나인 행동 인식에 입력으로 사용될 수 있다. 이에 따라 우수한 행동 구간 검출은 행동 인식 성능에 향상에 중요하다.

최근 Temporal Action Proposal Generation 방법들 중 하나인 SRG[2]는 단편 관련성 (Snippet

※ 이 논문은 2020년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (No.B0101-15-0266, 실시간 대규모 영상 데이터 이해·예측을 위한 고성능 비주얼 디스커버리 플랫폼 개발 및 No.2020-0-00004, 장기 시각 메모리 네트워크 기반의 예지형 시각지능 핵심기술 개발).

◆ First Author : (ORCID:0000-0001-7794-5377)School of Electrical Engineering, Korea Advanced Institute of Science and Technology, hj.eun@kaist.ac.kr, 학생(박사), 학생회원

° Corresponding Author : (ORCID: 0000-0003-3145-6732)Department of Electrical Engineering, Hanbat National University, peterjung@hanbat.ac.kr, 부교수, 정회원

* (ORCID:0000-0002-6616-824X, 0000-0002-4878-4129)Electronics and Telecommunications Research Institute (ETRI), {jymoon, jongyoul}@etri.re.kr, 책임연구원

** (ORCID:0000-0001-9323-8488) School of Electrical Engineering, Korea Advanced Institute of Science and Technology, changick@kaist.ac.kr, 정교수

논문번호 : 202004-078-A-LU, Received April 7, 2020; Revised May 1, 2020; Accepted May 8, 2020

Relatedness)를 제안하며 높은 성능 향상을 보여주었다. SRG는 Temporal Convolutional Networks (TCNs) 기반의 TIGN과 TIEN으로 이루어져 있으며, TIGN은 시간적 구간을 생성하고, 생성한 TIEN은 시간적 구간을 평가하여 행동 구간을 정의한다.

본 논문에서는 SRG에서 사용하는 TCN 기반의 TIGN이 아닌 Long Short-Term Memory (LSTM) 네트워크[6] 사용하여 단편 관련성을 학습한다. 그림 1은 제안하는 방법과 이전 방법들에 대한 비교를 보여준다. LSTM은 시간적 순서에 따라 시퀀스를 나누어 입력으로 사용한다. 이는 시간적으로 연속하여 일어나는 행동 모델링에 적합하다. 시간적 행동 검출 벤치마크 데이터 셋인 THUMOS-14[5]에 대한 실험을 통해 제안하는 방법의 우수성 보여주고 다양한 측면에서 분석을 수행하여 LSTM을 이용한 단편 관련성 학습의 가능성을 확인한다.

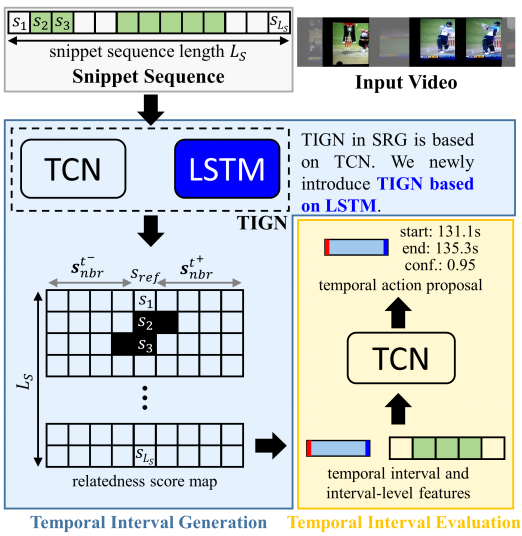

그림 1. 제안하는 방법과 SRG[2] 비교
Fig. 1. Comparison between the proposed method and SRG[2].

Ⅱ. 관련 연구

그림 2는 SRG[2]에서 제안하는 단편 관련성 점수 지도 생성 예를 보여준다. 단편 관련성은 단편 s가 특정 기준 행동에 포함 여부를 나타낸다. 여기서 기준 행동은 기준 단편 s_{ref}가 속하고 있는 행동으로 정의한다. 비디오의 모든 단편들이 기준 단편으로 정의가 되며 이를 통해 단편 관련성 지도를 생성할 수 있다.

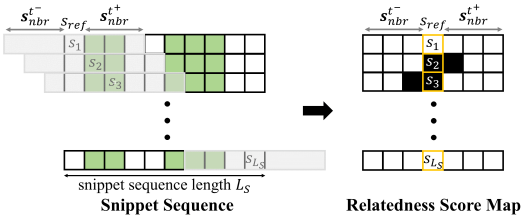

그림 2. 단편 관련성 점수 지도 생성 예
Fig. 2. Example of snippet relatedness score map.

단편 관련성 지도에서 s_{nbr}은 이웃 단편으로 기준 단편이 바뀔 때마다 비디오의 모든 단편들을 고려하는 것이 아닌 인접한 단편들만 고려하며 그 수는 최고 행동 길이로 정의한다. BSN[3]은 2단계로 이루어져 있다. 첫 번째 단계에서는 단편들의 시작, 끝, 행동 점수를 예측한다. 이 세 가지 중 시작과 끝의 점수가 높은 단편의 위치를 이용해 행동 구간을 정의하고, 두 번째 단계에서 해당 행동 구간들에 대한 행동 점수를 평가하여 행동 구간을 정의한다. MGG[4] 역시 두 개의 네트워크로 이루어져 있다. 하나의 네트워크는 임의의 행동 구간을 정의하여 해당 구간의 행동 점수를 평가하고, 다른 하나의 네트워크는 단편들의 행동 점수를 평가하여 행동 구간을 정의한다. 최종 행동 구간은 이 두 가지 결과를 합쳐서 얻게 된다.

Ⅲ. 제안하는 방법

그림 3은 단편 관련성 지도 생성을 위해 두 가지 LSTM 네트워크를 나타낸다. 첫 번째는 단방향 LSTM을 사용하여 네트워크를 구성한다. 그림에서와 같이 단편 특징 시퀀스는 각 LSTM의 입력으로 사용

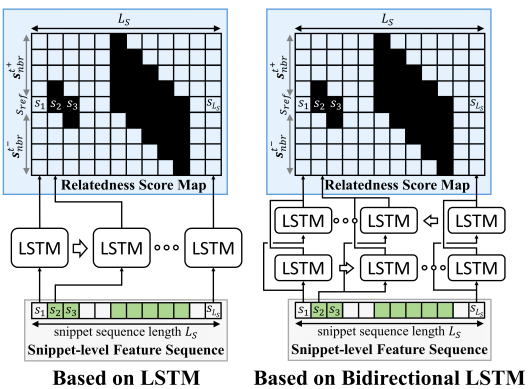

그림 3. 단편 관계성 생성을 위한 제안하는 LSTM 네트워크
Fig. 3. The proposed LSTM networks for relatedness score map generation.

되며 각 입력 단편이 기준 단편으로 정의된다. 이에 따라 각 LSTM은 단편 관계성 지도의 각 열을 평가한다. 단방향 LSTM의 경우 뒤 시간 단편의 LSTM 셀의 경우 앞 시간 단편의 정보를 활용하게 되지만 앞 시간 단편의 LSTM은 뒤 시간 단편의 정보없이 평가가 이루어진다.

이러한 문제를 해결하고자 두 번째 네트워크는 양방향 LSTM을 사용한다. 즉, 그림에서와 같이 과거의 정보가 미래 방향으로 축적될 뿐 아니라 미래 정보가 과거 방향으로 축적된다. 이에 따라 앞 시간 단편의 LSTM 셀 역시 뒤 시간 단편의 정보를 활용할 수 있다.

두 네트워크 모두 Binary Cross-Entropy Loss를 이용하여 단편 관계성 점수 지도를 학습하며 Loss는 아래와 같이 정의한다.

$$L_r = -\frac{1}{L_S L_{S_r}} \sum_{i=1}^{L_S} \sum_{j=1}^{L_{S_r}} (M_{r,ij} \log(O_{r,ij}) + (1 - M_{r,ij}) \log(1 - O_{r,ij})).$$

위 식에서 M은 단편 관계성 지도 레이블이며 O는 LSTM 네트워크에서 유추한 단편 관계성 지도이다. L_S와 L_{S_r}은 각각 비디오 길이와 단편 관계성 지도의 가로 길이를 나타낸다.

생성한 단편 관계성 지도의 각 열은 점수 시퀀스를 나타낸다. 각 시퀀스에서 높은 점수를 가지는 단편들을 그룹화하여 시간적 구간과 해당 구간에 대한 특징을 정의한다. 다음으로 정의한 시간적 구간에 대해 SRG의 TIEN을 이용하여 행동 신뢰 점수를 평가하고 시작 시간, 끝 시간, 행동 신뢰 점수를 통해 최종 행동 구간을 얻는다.

Ⅳ. 실 험

Dataset. THUMOS-14[5] 데이터 셋에 대해 제안하는 방법의 성능 평가를 수행한다. THUMOS-14 데이터 셋은 총 412개의 비정형 비디오와 20개 스포츠 행동에 대한 어노테이션을 포함하고 있다. 이 중 200개는 학습에 212개는 테스트에 사용한다.

Snippet-level features. LSTM 네트워크의 입력으로 사용하기 위해 비디오에 대해 단편 단위 특징을 추출하며 [2-4]에서 사용한 동일한 특징 추출기를 사용한다.

Evaluation metrics. 이전 방법들과 마찬가지로 성능 평가를 위해 행동 구간의 수를 변화시켜가며 평균 리콜 (Recall) 값을 측정한다. 평균 리콜은 tIoU를 0.5에서 1.0에 대해 계산된다.

표 1. THUMOS-14 데이터 셋 [5]에 대한 평균 행동 구간 수 대비 평균 리콜 성능 비교
Table 1. Performance comparison in terms of AR@AN on THUMOS-14 [5].

Method	@50	@100	@200	@500
BSN[3]	37.46	46.06	53.21	60.64
MGG[4]	39.93	47.75	56.71	61.36
SRG[2]	**42.19**	**49.72**	**56.71**	**63.78**
Ours (LSTM)	40.63	48.32	53.88	60.57
Ours (BiLSTM)	**41.34**	**49.14**	**55.60**	**63.11**

Performance comparison. 표 1은 이전 방법들과 제안하는 방법의 성능 비교를 나타낸다. 단방향 LSTM을 이용한 네트워크와 양방향 LSTM을 이용한 네트워크에 대해 성능을 측정하였다. SRG를 제외한 최신 방법들 대비 제안하는 방법이 모든 평균

행동 구간 수에서 가장 높은 리콜 성능을 보여준다.

빨간색은 가장 높은 성능을 파란색은 두 번째로 높은 성능을 나타낸다. 제안하는 방법이 TCN 기반의 SRG보다는 낮지만 유사한 성능을 보여주며, 이는 LSTM 기반의 TIGN의 유효성을 증명한다. 또한 단방향 LSTM보다 양방향 LSTM의 성능 차이를 통해 양방향 LSTM이 단편 관계성을 모델링 하는데 더 적합하다는 것을 알 수 있다.

Qualitative Evaluation. 그림 4는 제안하는 방법을 이용하여 생성한 시간적 행동 구간을 시각화한 결과이다. 실제 행동 구간 대비 시작과 끝 시간을 잘 평가하며, 높은 행동 신뢰 점수를 보여주고 있다.

Analysis. 그림 5는 TCN, 단방향 LSTM, 양방향 LSTM 기반의 네트워크를 사용하여 생성한 단편 관련성 점수 지도를 보여준다. TCN의 경우 SRG의

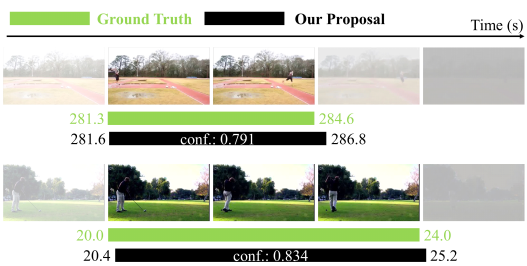

그림 4. 제안하는 방법을 통해 생성된 행동 구간에 대한 정성적 평가
Fig. 4. Qualitative evaluation of temporal action proposals generated from the proposed method.

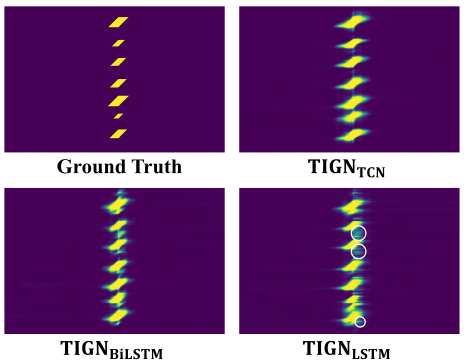

그림 5. 단편 관련성 점수 지도에 대한 정성적 평가
Fig. 5. Qualitative evaluation of relatedness score maps.

TIGN으로부터 얻어진 결과이다. 단방향 LSTM의 경우 기준 단편에 대한 점수 시퀀스의 경우 뒤 시간 단편의 정보를 사용할 수 없기 때문에 그림에서와 같이 기준 단편보다 뒤 시간 단편의 점수에 있어서 추정이 올바르지 않은 것을 확인할 수 있다. 양방향 LSTM을 사용한 경우에는 단방향 LSTM을 사용한 결과보다 훨씬 좋은 추정 결과를 보여주며, 단방향 LSTM 대비 미래 단편에서 나은 성능을 보임을 확인 할 수 있다. 하지만 TCN 기반의 TIGN과 비교 하였을 때 여전히 미래 단편에서 False Positive가 발생함을 보여준다. 단편 관련성 점수 지도에 대한 정량적 평가를 위해 Mean Absolute Error (MAE)를 측정하였다. Ground Truth 대비 TIGN$_{TCN}$, TIGN$_{BiLSTM}$, TIGN$_{LSTM}$은 각각 0.022, 0.032, 0.041의 MAE를 나타내었다.

V. 결 론

본 논문에서는 SRG[2]의 TCN 기반의 TIGN을 단방향, 양방향 LSTM 기반으로 새롭게 제안하여 시간적 행동 구간 생성을 수행하였다. 실험을 통해 제안하는 방법이 최근 시간적 행동 구간 생성 방법인 BSN[3], MGG[4] 대비 높은 성능을 SRG와는 유사한 성능을 나타냄을 보여주었다. 또한 기존 SRG의 TCN 기반의 TIGN과 비교 및 분석을 수행하였으며, LSTM 기반의 TIGN의 유효성을 확인하였다. 본 연구는 다양한 시퀀스 모델링 기반의 시간적 행동 구간 생성 연구에 하나의 베이스 모델로 도움이 될 것으로 판단된다.

References

[1] J. Park and S. Kwak, "Detection of crowd escape behavior in surveillance video," *J. KICS*, vol. 36, no. 8, pp. 731-737, Aug. 2014.

[2] H. Eun, et al., "SRG: Snippet relatedness-based temporal action proposal generator," *IEEE TCSVT*, DOI:10.1109/TCSVT.2019.2953187.

[3] T. Lin, X. Zhao, H. Su, C. Wang, and M. Yang, "Bsn: Boundary sensitive network for temporal action proposal generation," in *Proc. Eur. Conf. Comput. Vis.*, pp. 3-19, Sep. 2018.

[4] Y. Liu, L. Ma, Y. Zhang, W. Liu, and S.-F. Chang, "Multi-granularity generator for temporal action proposal," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 3604-3613, Jun. 2019.

[5] Y. G. Jiang, et al., "*Thumos challenge: Action recognition with a large number of classes*," 2014, http://crcv.ucf.edu/THUMOS14/.

[6] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735-1780, Dec. 1997.