특집논문 (Special Paper)
방송공학회논문지 제30권 제6호, 2025년 11월 (JBE Vol.30, No.6, November 2025)
https://doi.org/10.5909/JBE.2025.30.6.899
ISSN 2287-9137 (Online) ISSN 1226-7953 (Print)

시각-언어 모델이 생성한 설명문의 효과적인 융합을 통한 비디오 구간 검색 및 하이라이트 감지 기법

이 유 은[a], 김 정 욱[b)‡]

Video Moment Retrieval and Highlight Detection via Effective Fusion of Captions Generated by Vision-Language Models

YuEun Lee[a)] and Jung Uk Kim[b)‡]

요 약

텍스트 쿼리를 활용한 비디오 순간 검색(Moment Retrieval, MR)과 하이라이트 감지(Highlight Detection, HD)는 방대한 비디오 데이터에서 사용자의 관심 구간이나 중요한 장면을 찾는 것을 목표로 한다. 기존 연구들은 주로 시각 정보에 기반해 텍스트 쿼리와 비디오 간의 대응을 학습하였으나, 이는 장면의 맥락적 이해가 부족하다는 한계가 있다. 본 논문에서는 이러한 한계를 극복하기 위해 시각-언어 모델(Vision-Language Models, VLMs)로 각 비디오 클립의 캡션을 생성하고, 이를 텍스트 쿼리 및 비디오와 통합하는 새로운 프레임워크를 제안한다. 쿼리 기반 attention과 cross-attention 메커니즘을 통해 비디오와 캡션을 정제된 표현으로 통합하고, 이를 기반으로 MR과 HD를 동시에 수행한다. 실험 결과, 제안 방법은 기존 방법 대비 우수한 성능을 보이며 캡션 기반 장면 이해가 두 작업의 성능 향상에 효과적임을 입증하였다.

Abstract

Video moment retrieval (MR) and Highlight detection (HD) using text queries aim to identify user-interested segments or important scenes from massive video data. Existing research primarily relies on visual information to learn correspondences between text queries and videos, but this lacks contextual understanding. To overcome these limitations, this paper proposed a novel framework that generates captions for each video clip using Vision-Language Models (VLMs) and integrates these captions with text queries and videos. Using query-based attention and cross-attention mechanisms, the video and captions are integrated into a refined representation, enabling simultaneous MR and HD. Experimental results demonstrate that the proposed method outperforms existing methods, demonstrating the effectiveness of caption-based scene understanding in both tasks.

Keyword : Moment Retrieval, Highlight Detection, Vision-Language Model, Caption Generation

Ⅰ. 서 론

최근 디지털 기기와 인터넷 플랫폼의 확장으로 비디오 콘텐츠가 기하급수적으로 증가하고 있다. 이러한 방대한 양의 비디오 콘텐츠로부터 사용자가 원하는 정보를 효율적으로 추출하는 것은 점점 더 어려워지고 있으며, 이를 해결하기 위해 자연어 텍스트 쿼리를 기반으로 관련 장면을 자동으로 찾는 연구가 활발히 진행되고 있다. 대표적이 두 가지 연구는 비디오 내에서 자연어 쿼리와 의미적으로 관련된 특정 순간을 찾는 비디오 순간 검색(Moment Retrieval, MR)과 비디오의 가장 중요한 클립을 식별하는 하이라이트 감지(Highlight Detection, HD)이다.

기존에는 MR과 HD가 독립적으로 다루어졌지만, 두 작업 모두 사용자 의도에 부합하는 핵심 장면을 추출한다는 공통점을 지니고 있어, 최근에는 이를 통합적으로 해결하려는 시도가 이루어지고 있다. QVHighlights 데이터셋과 Moment-DETR[1]의 도입으로 두 작업을 동시에 처리할 수 있는 기반이 마련되었다. UMT[2]는 청각 모달리티를 통합한 구조를 제안하였으며 QD-DETR[3]은 비디오-텍스트 쌍 간의 부정적 관계를 학습하는 손실 함수를 설계하였다. 또한 최신 모델인 TR-DETR[4]은 MR과 HD 간의 상호작용을 강조하며 성능을 향상시켰다.

최근 연구들은 순간 검색 및 하이라이트 감지 작업의 유사점과 차이점을 고려하여 상당한 개선을 이루었지만, 대부분 비디오의 시각적 정보에 과도하게 의존하고 있어 장면의 맥락적 이해나 비디오에 내재된 의미적 정보는 충분히 활용하지 못하는 한계가 있다. 실제로 비디오 클립에는 단순히 시각적 요소 외에도, 객체 간의 공간적 상호작용, 동작의 흐름, 장면의 시간적 변화와 같은 장면 이해 정보가 포함되어 있으며, 이러한 요소들은 쿼리와의 정밀한 의미 정렬을 위해 매우 중요하다. 이러한 추가적인 정보를 분석함으로써 비디오 장면에 대한 심층적인 이해를 얻고 텍스트 쿼리와 가장 관련 있는 비디오 클립을 찾을 수 있다.

이에 본 논문에서는 시각-언어 모델(Vision-Language Models, VLMs)을 활용하여 각 비디오 클립에 대한 자연어 기반 설명문인 캡션을 생성하고 이를 기존 비디오 및 텍스트 쿼리와 효과적으로 통합하는 새로운 프레임워크를 제안한다. 캡션은 해당 클립의 주요 시각적 요소와 의미를 요약하여 표현함으로써 단순한 픽셀 기반 정보가 아닌 고수준의 의미적 정보를 제공한다. 본 연구에서는 이러한 캡션을 쿼리 기반 attention과 cross-attention 메커니즘을 통해 비디오 및 쿼리와 결합함으로써 시각 정보와 언어 정보 간의 의미적 연결을 정교하게 구축하고자 한다. 이를 통해 모델이 보다 정밀하게 쿼리의 의도를 반영할 수 있다. 궁극적으로, 본 연구는 캡션으로부터 유도된 의미적 단서를 통해 기존 방식보다 더욱 정확한 순간 검색 및 하이라이트 감지 성능을 달성하고자 한다.

Ⅱ. 본 론

1. 시각-언어 모델을 이용한 설명문 생성

최신 시각-언어 모델인 InternVL2[5]를 채택하여 개별 비디오 클립에 대한 설명문인 캡션을 추출한다. 캡션은 각 클립에 대해 자세한 의미적 설명을 제공한다. 모델에 제공된 프롬프트는 다음과 같다. “Please describe the image in one sentence.”. 설명을 단일 문장으로 제한함으로써, 각 비디오 클립의 핵심 시각적 요소에 집중하도록 안내하면서 불필요한 세부 정보를 피한다. 결과적으로, 캡션의 사용은 비디오 클립의 시각적 정보뿐만 아니라 텍스트 정보도 효과적으로 활용하여 구간 검색 및 하이라이트 감지를 위한 풍부한 맥락 정보를 제공한다.

a) 경희대학교 컴퓨터공학과(The department of Computer Science and Engineering, Kyung Hee University)
b) 경희대학교 컴퓨터공학부(School of Computing, Kyung Hee University)
‡ Corresponding Author : 김정욱(Jung Uk Kim)
E-mail: ju.kim@khu.ac.kr
Tel: +82-31-201-3768
ORCID: https://orcid.org/0000-0003-4533-4875
※ 이 논문의 결과 중 일부는 한국방송·미디어공학회 2025년 하계학술대회에서 발표한 바 있음
※ This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No.RS-2022-00155911, Artificial Intelligence Convergence Innovation Human Resources Development (Kyung Hee University))
· Manuscript September 8, 2025; Revised October 17, 2025; Accepted October 20, 2025.

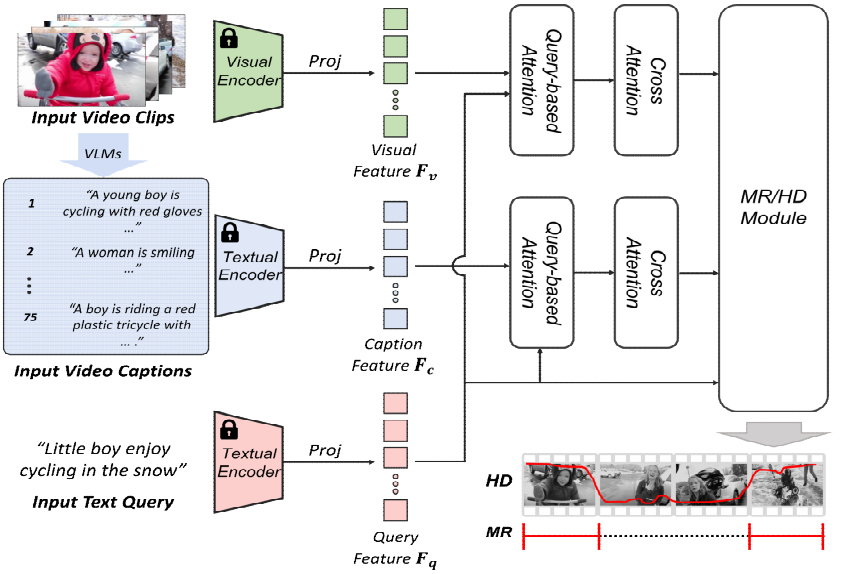

그림 1. 제안하는 방법의 전체 구조
Fig. 1. Overall architecture of the proposed method

2. 전체 프레임워크

그림 1은 제안하는 방법의 전체 구조를 나타낸다. 최신 연구인 TR-DETR[4]을 baseline 모델로 설정하고 시각-언어 모델로 생성한 캡션을 통합한다. 입력은 L_v개의 클립이 있는 비디오와 L_q개의 단어를 가진 텍스트 쿼리로 구성된다. 각 입력은 인코더를 통과한 후 3-layer feed-forward 네트워크를 지나 비디오 특징 F_v와 텍스트 쿼리 특징 F_q를 생성한다. 또한 각 비디오 클립에 대해 시각-언어 모델로 생성된 캡션은 같은 과정을 거쳐 캡션 특징 F_c를 생성한다.

먼저, 비디오 및 캡션과 텍스트 쿼리의 의미 정렬을 향상시키기 위해 우리는 쿼리 기반 attention 메커니즘을 진행한다. 구체적으로, 쿼리-비디오 유사도 행렬 및 쿼리-캡션 유사도 행렬을 계산한 후 각 유사도 행렬에 softmax 정규화를 취해 attention 가중치를 생성한다. 이후 가중치를 텍스트 쿼리에 곱한 후 비디오 및 캡션 특징과 접합(concatenation)함으로써 쿼리로부터 유도된 비디오 특징 F_{vq} 및 캡션 특징 F_{cq}를 생성한다.

이후 두 쿼리 기반 특징을 서로 교차적으로 통합하기 위해 cross-attention을 적용한다. F_{vq}에 대해서는 query를 F_{vq}, key와 value를 F_{cq}로 설정하고, F_{cq}에 대해서는 query를 F_{cq}, key와 value를 F_{vq}로 설정한다. 이를 통해 비디오에 대한 시각적 정보인 비디오 특징과 텍스트 정보인 캡션 특징을 상호 보완적으로 통합하고 장면 맥락을 보다 심층적으로 해석할 수 있다. 캡션과 비디오는 최종적으로 접합되어 하나의 정제된 비디오 특징이 된다. 즉, 쿼리 기반 attention을 통해 텍스트 쿼리 모달리티와의 정렬을 강화하고 cross-attention을 통해 비디오와 캡션 모달리티의 정보를 병합하여 의미적 수준의 연관성을 향상시키는 것이다.

마지막으로, 순간 검색 및 하이라이트 감지를 위해 baseline인 TR-DETR[4]과 동일한 예측 head를 사용한다. 예측 head는 트랜스포머 인코더와 디코더로 구성되며 텍스트 쿼리 특징과 정제된 비디오 특징이 인코더의 입력으로 사용된다.

3. 손실 함수

학습을 위해 순간 검색의 L_{mr}, 하이라이트 감지의 L_{hd}, baseline 모델의 L_{local}과 L_{global} 손실을 결합하여 사용한다.

또한 비디오 클립에 대한 캡션과 해당 클립의 시각적 표현 간의 의미적 일치를 학습하기 위한 손실 함수인 L_{align}을 추가로 고안하였다. 이 손실 함수는 다음과 같이 계산된다.

$$L_{align} = -\frac{1}{B}\sum_{k=1}^{B}\log\frac{\sum_{j=1}^{L_v}\exp(sim(F_{v_{kj}}, F_{c_{kj}}))}{\sum_{i=1}^{B}\sum_{j=1}^{L_v}\exp(sim(F_{v_{ij}}, F_{c_{kj}}))} \quad (1)$$

이는 contrastive 손실 함수로, 동일한 캡션-비디오 클립 쌍의 유사도를 높이고 다른 쌍의 유사도는 낮추도록 학습되며 결과적으로 장면 수준에서 의미 정렬이 강화된다.

Ⅲ. 실험 및 결과

1. Dataset

비디오 순간 검색 및 하이라이트 감지를 동시에 수행하기 위해 QVHighlights dataset[1]을 사용하였다. 총 10,148개의 유튜브 기반 비디오 클립으로 구성되어 있으며 각 비디오는 텍스트 쿼리와 쿼리에 대응하는 정답 구간을 포함한다. 순간 검색 및 하이라이트 감지 작업을 동시에 수행하도록 설계된 대규모 비디오-텍스트 데이터셋이기 때문에 두 가지 작업에 대한 label이 모두 제공된다.

2. 실험 설정 및 평가지표

비디오 특징을 추출하기 위해 SlowFast[6]와 CLIP[7] 인코더를 사용하였고 텍스트 특징을 추출하기 위해 CLIP[7] 인코더를 사용하였다. 한 개의 NVIDIA RTX 3090 GPU를 사용하였으며 AdamW optimizer를 사용하였다. Cross-attention은 8개 head를 가진 multi-head attention이다. 학습을 위해 32의 batch size, 학습률 1e-4, epoch는 200으로 설정하였다.

평가에서는 이전 연구들[1-4]과 동일한 평가지표를 채택하였다. 순간 검색에 대해, Intersection over Union(IoU) 기준으로 임계값이 각각 0.5, 0.7인 Recall@1(R1@0.5, R1@0.7)을 사용하여 예측된 구간이 정답 구간과 일정 수준 이상 겹치는지를 평가하였다. 또한, 다양한 IoU 임계값(0.5-0.95, 간격 0.05)에 대해 평균 정밀도(mean Average Precision, mAP)를 계산하여 전반적인 성능을 측정하였다. 하이라이트 감지에 대해서는, 각 클립의 중요도 예측 정확도를 측정하는 mAP(mean Average Precision)와 가장 점수가 높은 클립이 실제 하이라이트인지 여부를 평가하는 HIT@1 지표를 사용하였다. 이를 통해 모델이 텍스트 쿼리와 의미적으로 잘 부합하는 비디오 구간을 얼마나 정확히 예측하는지를 정량적으로 분석하였다.

3. 성능 비교

표 1은 제안하는 방법의 성능을 기존 MR 및 HD 모델인 M-DETR[1], QD-DETR[3], UniVTG[8], TR-DETR[4]과 비교한 결과이다. 모든 평가지표에 대해 기존 모델보다 더욱 우수한 성능을 보임을 확인할 수 있다. 특히, HD의 성능이 급격히 향상했는데, 이를 통해 비디오 클립을 설명하는 텍스트 정보인 캡션이 비디오 내의 가장 중요한 클립을 식별하는 데 중요한 역할을 한다는 것을 알 수 있다.

표 1. 기존 연구들과 제안한 방법의 실험 결과 비교
Table 1. Comparison of experimental results of the proposed method with existing studies

Method	MR					HD	
	R1		mAP			≥Very Good	
	@0.5	@0.7	@0.5	@0.75	Avg.	mAP	HIT@1
M-DETR[1]	52.89	33.02	54.82	29.40	30.73	35.69	55.60
QD-DETR[3]	62.40	44.98	62.52	39.88	39.86	38.94	62.40
UniVTG[8]	58.86	40.86	57.60	35.59	35.47	38.20	60.96
TR-DETR[4]	64.66	48.96	63.98	43.73	42.62	39.91	63.42
Proposed Method	**69.16**	**53.81**	**67.49**	**47.32**	**46.81**	**42.86**	**69.61**

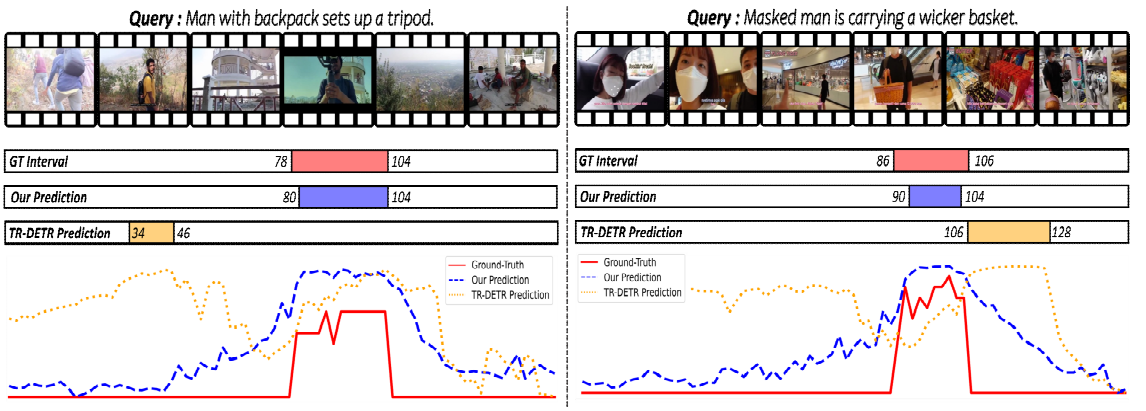

그림 2. MR 및 HD의 예측 시각화 결과
Fig. 2. Visualization results for MR and HD

4. 시각화 결과

그림 2는 예측 결과를 baseline 모델인 TR-DETR[4]과 비교한 시각화 그림이다. 동일한 입력 쿼리에 대해 제안한 모델이 기준 모델보다 GT 구간과 더 높은 일치를 보임을 알 수 있다. 시각화 결과는 우리의 방법이 기존 방법보다 순간 검색 및 하이라이트 감지에 더 좋은 효과를 보임을 증명한다.

Ⅳ. 결 론

본 논문에서는 비디오 순간 검색(MR)과 하이라이트 감지(HD) 작업에서 비디오의 시각적 정보에만 의존하는 기존 방법의 한계를 극복하기 위해, 시각-언어 모델(VLMs)을 활용하여 각 비디오 클립에 대한 의미적 캡션을 생성하고 이를 통합하는 새로운 프레임워크를 제안하였다. 제안한 모델은 텍스트 쿼리를 중심으로 비디오와 캡션 간의 의미 정렬을 강화하는 쿼리 기반 attention과, 두 모달리티 간 상호 보완적인 정보를 통합하는 cross-attention 메커니즘을 통해 정제된 비디오 표현을 학습한다. 또한, 캡션과 비디오 간 의미적 일치를 학습하는 추가 손실 함수를 도입하여 장면 수준의 정합성을 높였다. 실험 결과, 제안 모델은 기존 baseline인 TR-DETR 대비 모든 평가지표에서 우수한 성능을 보였으며, 정량적 지표뿐만 아니라 정성적 시각화에서도 높은 정확도를 확인할 수 있다. 이를 통해, 캡션 기반 장면 맥락 정보가 MR 및 HD 성능 향상에 실질적인 기여를 함을 입증하였으며 향후 연구 및 응용 분야에서 본 접근 방식의 잠재력을 보여준다.

참 고 문 헌 (References)

[1] Lei, Jie, Tamara L. Berg, and Mohit Bansal, “Detecting moments and highlights in videos via natural language queries,” *Advances in Neural Information Processing Systems*, pp. 11846-11858, 2021.

[2] Liu, Ye, et al, “Umt: Unified multi-modal transformers for joint video moment retrieval and highlight detection,” *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3042-3051, 2022.
doi: https://doi.org/10.1109/CVPR52688.2022.00305

[3] Moon, WonJun, et al, “Query-dependent video representation for moment retrieval and highlight detection,” *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23023-23033, 2023.
doi: https://doi.org/10.1109/CVPR52729.2023.02205

[4] Sun, Hao, et al, “Tr-detr: Task-reciprocal transformer for joint moment retrieval and highlight detection,” *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol.38, No.5, pp. 4998-5007, 2024.
doi: https://doi.org/10.1609/aaai.v38i5.28304

[5] Chen, Zhe, et al, “Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks,” *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24185-24198, 2024.
doi: https://doi.org/10.1109/CVPR52733.2024.02283

[6] Feichtenhofer, Christoph, et al, “Slowfast networks for video recognition,” *Proceedings of the IEEE/CVF International Conference*

on Computer Vision, pp. 6202-6211, 2019.
doi: https://doi.org/10.1109/ICCV.2019.00630

[7] Radford, Alec, et al, “Learning transferable visual models from natural language supervision,” *International Conference on Machine Learning*, PmLR, pp. 8748-8763, 2021.

[8] Lin, Kevin Qinghong, et al, “Univtg: Towards unified video-language temporal grounding,” *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2794-2804, 2023.
doi: https://doi.org/10.1109/ICCV51070.2023.00262

저 자 소 개

이 유 은

- 2024년 : 경희대학교 생체의공학과 및 전자공학과 졸업(학사)
- 2024년 ~ 현재 : 경희대학교 컴퓨터공학과 석박통합과정
- ORCID : https://orcid.org/0009-0002-9364-5731
- 주관심분야 : 컴퓨터비전, 딥러닝, 멀티모달

김 정 욱

- 2016년 : 아주대학교 전자공학과 졸업(학사)
- 2018년 : 한국과학기술원(KAIST) 졸업(석사)
- 2022년 : 한국과학기술원(KAIST) 졸업(박사)
- 2022년 ~ 현재 : 경희대학교 컴퓨터공학부 조교수
- ORCID : https://orcid.org/0000-0003-4533-4875
- 주관심분야 : 딥러닝, 시각 인공지능, 멀티모달, 3D 객체 검출