논문 2022-17-41

텍스트-비디오 검색 모델에서의 캡션을 활용한 비디오 특성 대체 방안 연구

(A Study on the Alternative Method of Video Characteristics Using Captioning in Text-Video Retrieval Model)

이 동 훈† , 허 찬† , 박 혜 영*, 박 상 효*
(Dong-hun Lee, Chan Hur, Hyeyoung Park, Sang-hyo Park)

Abstract : In this paper, we propose a method that performs a text-video retrieval model by replacing video properties using captions. In general, the exisiting embedding-based models consist of both joint embedding space construction and the CNN-based video encoding process, which requires a lot of computation in the training as well as the inference process. To overcome this problem, we introduce a video-captioning module to replace the visual property of video with captions generated by the video-captioning module. To be specific, we adopt the caption generator that converts candidate videos into captions in the inference process, thereby enabling direct comparison between the text given as a query and candidate videos without joint embedding space. Through the experiment, the proposed model successfully reduces the amount of computation and inference time by skipping the visual processing process and joint embedding space construction on two benchmark dataset, MSR-VTT and VATEX.

Keywords : Multimodal Deep Learning, Video-Captioning, Text-Video Retrieval

Ⅰ. 서 론

텍스트-비디오 검색은 텍스트가 입력으로 주어질 때 데이터베이스안의 후보 비디오들에서 의미적으로 일치하는 비디오를 찾아내는 과제이다. 그림 1은 텍스트-비디오 검색에 대한 예시로 왼쪽의 텍스트 쿼리를 입력으로 받았을 때 그에 맞는 비디오를 찾는 텍스트-비디오 검색 결과를 시각적으로 설명하고 있다. 최근 급격히 성장한 Netflix, YouTube 등의 비디오 스트리밍 시장과 더불어 많은 주목을 받고 다양한 연구가 이루어지고 있는 분야이다. 이러한 과제를 해결하기 위해 많은 연구들이 제안되고 있는데 가장 널리 사용되는 방법은 임베딩 기반 접근방법이다. 임베딩 기반 접근 방법은 비디오와 텍스트의 특징을 추출하여 같은 공간상에서 매핑하는 방법으로 비디오와 텍스트라는 종류가 다른 두 데이터를 특징 추출 모듈을 통해 특징을 추출한 후 공통되는 공간상에 임베딩시킨다.

기존의 공통 임베딩 기반 연구에서는 임베딩 하고자 하는 비디오와 텍스트를 각각 별도의 특징 추출 모듈을 이용하여 특징을 추출한다. 예를 들어, 비디오의 경우 CNN 계열 모델

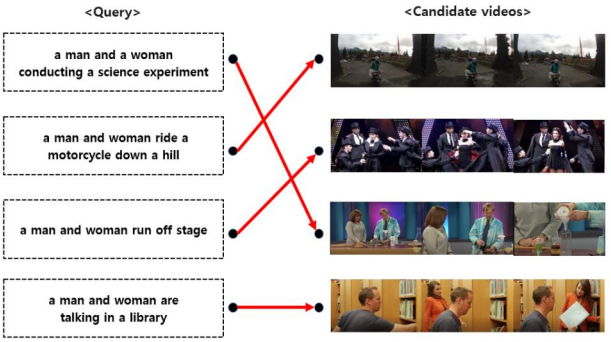


그림 1. 텍스트-비디오 검색에 대한 예시
Fig. 1. Examples of text-video retrieval

을 인코더로 사용하여 프레임 단위로 이미지의 시각적인 특징벡터를 추출하고, 텍스트의 경우에는 순차정보를 얻을 수 있는 RNN 계열 인코더를 기반으로 하여 문장의 특징벡터를 추출한다. 이러한 특징벡터들을 임베딩 공간상에서 맵핑하는 경우 각각 텍스트와 이미지라는 다른 데이터의 형태를 공통 공간에 임베딩하기 때문에 두 데이터의 분포 차이에서 오는 격차문제 [1]가 발생하게 된다. 또한 이와 같은 구조를 이용하는 선행 연구들 [2-4]에서는 텍스트 인코더 대비 비디오의 시각정보 처리 과정에서의 연산량이 급격히 증가하게 된다. 비디오의 경우 프레임 단위별로 이미지를 기반으로 한 시각적인 정보를 담고 있을 뿐만 아니라 프레임의 연속적인 내용을 담고 있는 시간적인 정보도 담고 있어 인코딩시 많은 연산량을 요구한다. 그로 인해 비디오 시각 정보

† These authors contributed equally to this work.
*Corresponding Authors (hypark@knu.ac.kr; s.park@knu.ac.kr)
Received: Oct. 20, 2022, Revised: Nov. 11, 2022, Accepted: Nov. 26, 2022.
D. Lee: Kyungpook National University (Undergraduate Student)
C. Hur: Kyungpook National University (PhD Student)
H. Park: Kyungpook National University (Prof.)
S. Park: Kyungpook National University (Asst. Prof.)
※ 본 논문은 2020년도 정부 (교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임 (No. 2020R1I1A3072227).
※이 논문의 연구 결과 중 일부는 한국통신학회 “제3회 한국 인공지능 학술대회”에서 발표한 바 있음.

ⓒ IEMEK J. Embed. Sys. Appl. 2022 Dec. 17(6) 347-353
ISSN : 1975-5066
http://dx.doi.org/10.14372/IEMEK.2022.17.6.347

를 처리하기 위해 여러 복잡한 과정이 더해지게 되어 전체 모델의 크기가 커지게 된다. 우리는 이러한 비디오 시각처리 부분에서 생기게 되는 많은 연산 과정을 줄이고자 비디오의 특성을 생성모듈을 통한 캡션으로 대체하여 텍스트-비디오 검색을 진행하는 모델을 제안한다.

제안 모델은 비디오의 특성을 비디오의 정보를 표현하는 캡션 (텍스트)으로 변환한 뒤 이 캡션과 입력으로 받는 쿼리 텍스트를 비교하여 검색을 하게 된다. 쿼리와 동일한 유형의 텍스트 데이터를 처리를 할 수 있게 되므로, 복잡한 비디오 시각처리 과정을 요구하지 않기 때문에 기존에 비디오와 텍스트를 임베딩할 때 각각의 특성의 차이에서 생기는 격차 문제를 줄일 수 있게 된다. 이를 통해 모델을 경량화 시킬 수 있으며 그로 인해 추론 단계에서 사용하는 연산량 및 연산시간도 줄어들게 된다. 제안 방법의 성능을 검증하기 위해서 MSR-VTT, VATEX 2개의 벤치마크 데이터셋에서 실험 및 분석을 진행하였다.

Ⅱ. 관련 연구

1. 텍스트-비디오 검색

텍스트-비디오 검색은 크게 개념기반, 공통 임베딩 공간 기반 방법으로 나뉘게 된다. 개념 기반 방법의 대표적인 예시는 Ad-hoc Video Search (AVS) [5]가 있다. AVS는 시각적 개념 분류기와 언어 규칙에 의존하며, 텍스트 쿼리와 특정 비디오 사이의 유사성을 개념 일치를 통해 계산을 한다. AVS의 경우 비디오와 텍스트 쿼리를 개념으로 표현해 주기 때문에 어느 정도 해석이 가능하다는 장점이 있지만, 비디오와 쿼리 둘 다를 통해 전체적인 관계가 매핑된 정보를 설명하는 것이 매우 어렵다는 단점도 있다.

이러한 문제를 해결하기 위해 가장 많이 사용되는 방법은 공통 임베딩 기반 방법 [2-4, 6]이다. 공통 임베딩 기반 방법은 비디오 및 텍스트 쿼리를 인코딩 한 후 공통의 잠재된 공간상에 위치시켜 학습 및 추론에 사용하는 방법이다. 많은 연구들이 이 과정에서 사전 훈련된 CNN 기반 모델을 이용한 정교한 비디오 인코딩 과정을 통해 비디오를 특정 프레임 단위로 나눠 이미지 레벨에 있는 시각적 특징들을 추출해 비디오 피쳐로 정의하는 인코딩된 결과물로 표현하였다. 초기 연구에는 비디오 프레임간의 시간 순서와 가중치를 고려하지 않고 평균 풀링 [2]을 통해 집계하는 방법을 사용하였으나 점차 비디오 프레임의 시간 정보를 명시적으로 모델링하기 위해 LSTM [4], GRU [5], 최댓값 풀링 [7, 8] 또는 멀티헤드-셀프 어텐션 [9] 등을 사용하여 시간 순서와 프레임간의 중요도를 모델링하였다.

상술한 많은 방법들은 비디오의 시각적인 정보와 시간적인 정보를 담기 위해 인코딩 과정에서 여러 모듈을 사용하게 된다. 이런 과정은 비디오의 길이에 따라 유동적으로 바뀔 뿐 아니라 많은 연산량을 가지게 되어 모델이 복잡해져 추론 과정에서조차 시간과 자원을 많이 소모하는 단점을 가진다. 우리가 제안하는 모델은 비디오의 특성을 캡션으로 대체하는 방법을 사용하기 때문에 입력으로 받는 쿼리 텍스트와 후보 비디오의 표현 형태가 동일한 형태를 가지게 된다. 이러한 같은 형태의 데이터의 이용은 비디오 시각 인코더를 생략하기 때문에 비디오 입력에 대한 연산량을 상당히 감소시켰다.

2. 비디오-캡셔닝

비디오-캡셔닝이란 비디오가 주어질 때 이를 언어적으로 설명하는 캡션 문장을 생성하는 과제이다. 최근 딥러닝을 활용한 자연어 처리와 컴퓨터 비전의 발전을 통해 많은 비디오 캡셔닝 모델들이 제안되고 있으며 많은 연구 [10-12]가 인코더-디코더 기반의 구조를 사용하고 있다. 인코더-디코더 기반의 방법은 입력으로 받은 텍스트와 비교 대상이 되는 비디오를 각각 별도의 인코더에 넣어서 인코딩 과정을 거친 뒤 비교를 하여 나온 결과물을 디코더에 넣어서 보여주는 방법이다.

인코더에서 비디오의 시각적인 정보를 인코딩하기 위해 주로 CNN 계열 모델들을 인코딩 모듈로 활용한다. CNN을 활용한 방법 [13]은 비디오를 프레임 단위로 나누어 특성을 추출한다. 이때 일부 모델 [14, 15]은 어텐션 메커니즘을 사용하여 각 프레임간의 연관성을 찾기도 하며, 오디오의 정보도 활용하는 방식도 있다. 또한 모든 프레임이 아닌 특정 프레임만을 채택해서 캡션을 만드는 방법도 있다. 디코더에서 텍스트의 특성을 추출하는 과정에는 주로 RNN [13] 계열 방법들을 사용한다. 문장에서의 단어의 순서를 결정하는데 있어 연속적인 데이터의 특징을 잘 추출하는 RNN 기반의 방법을 주로 사용하며, 최근에는 어텐션 기법을 사용하거나 Transformer 모델 [7]을 사용하여 비디오 캡션을 생성하는 방법들 [10, 16, 17] 도 존재한다.

본 논문에서는 Transformer를 사용한 인코더-디코더 구조의 비디오 캡션 모델 [10]을 이용하여 비디오의 특성으로 활용하였다. 우리는 이러한 비디오-캡셔닝의 결과로 생성된 캡션이 비디오의 내용과 의미적으로 같지만 텍스트로 표현한다는 점에 주목하였다. 다른 관점에서 비디오를 표현하는 생성된 캡션 정보를 이용하여 텍스트-비디오 검색 문제에서 비디오의 특징을 효율적으로 표현하는 방법을 제안하였다.

Ⅲ. 제안 방법

3.1. 특징 추출 과정

본 논문에서는 텍스트-비디오 검색에서 인코딩 과정을 위한 기본 모델로 듀얼 인코더 모델 [18]의 구조를 사용하였다. 그림 2 하단부에서 볼 수 있듯이, 텍스트가 입력으로 주어질 때, 특징 추출 모듈 $\phi(t)$는 3단계의 특징추출과정을 거치며 다음과 같이 표현된다.

$$\phi(t) = [L_1(t), L_2(L_1(t)), L_3(L_2(L_1(t)))]. \quad (1)$$

단, 여기서 t는 입력으로 받는 텍스트 쿼리, L_1, L_2, L_3,는

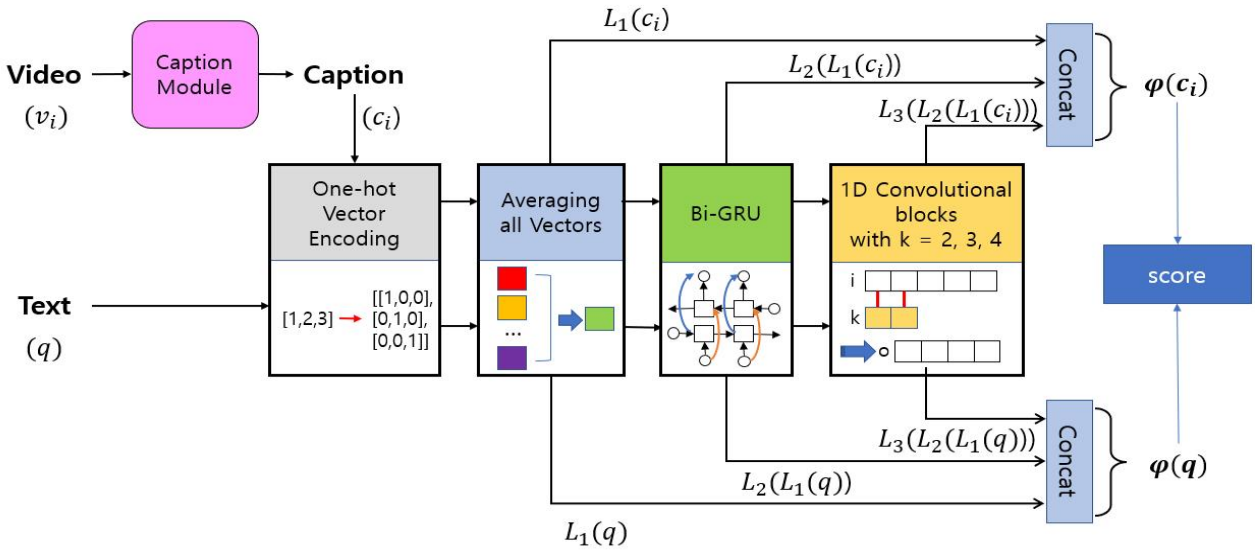


그림 2. 제안하는 텍스트-비디오 검색 모델
Fig. 2. Proposed text-video retrieval model

텍스트의 특징 추출과정 3단계를 말하며, $[\bullet, \bullet]$는 L_1, L_2, L_3의 결합한다는 것을 의미한다. 우선 입력으로 텍스트(t) 가 주어지면 이를 원-핫 벡터로 표현한다. 그러고 첫 번째 단계 (L_1)에서 이 벡터의 평균을 구해 텍스트를 전체적으로 표현하는 글로벌 인코딩 벡터 값을 얻는다. 다음, 두 번째 단계 (L_2)에서는 L_1을 통과한 글로벌 벡터를 bi-GRU를 통과시켜 텍스트의 순차적인 정보를 표현하는 벡터를 얻는다. 세 번째 단계 (L_3)에서는 L_2를 통과한 벡터를 1-D CNN을 통과시켜 텍스트의 지역정보를 반영하는 벡터를 얻는다. 최종적인 인코딩 결과로는, 이 세 단계에서 나온 벡터들을 결합하여 특징 벡터로 정의한다. 이러한 정교한 특징 추출 과정은 3.2장에서 도입한 비디오 특징을 캡션으로 대체한 생성 캡션들에 대해서도 같은 방법으로 진행되어진다.

3.2. 캡셔닝 모듈 선택

비디오 정보를 시각적으로 표현하는 방법 대신, 우리는 비디오의 정보를 생성 모델을 통해 얻어진 언어적 정보를 이용하여 표현하였다. 이러한 표현 방법은 입력 비디오가 쿼리와 같은 특성 (텍스트)으로 맞춰진다는 점에서 착안되었고 이를 표현 방법을 통해 얻은 장점을 3.3장에서 설명하도록 한다.

캡셔닝 생성 모델로는 UniVL [10]을 사용하였다. UniVL은 비디오 캡셔닝에 사용하는 인코더-디코더 기반의 사전학습 모델로 비디오와 텍스트를 이해하고 캡션을 생성하기 위해 4개의 모듈로 설계되었다. 4개의 모듈은 3개의 인코더와 1개의 디코더로 구성이 되어 있으며, 비디오와 텍스트를 각각 트랜스포머 인코더로 처리를 한 다음, 전처리한 비디오와 텍스트를 결합해 또 다른 트랜스포머를 기반해 만든 교차 인코더에 넣는다. 마지막으로 인코딩한 값을 디코더에 넣어 최종적인 캡션을 생성하게 된다. 실험에 사용하는 학습데이터셋으로 UniVL에 파인-튜닝해 얻어진 캡션 결과물을 듀얼 인코딩 모델의 비디오 특성으로 넣어 사용하였다.

3.3. 캡션정보를 이용한 추론 과정

3.2절에서 선택된 캡션 모듈을 통해 추론 과정에서 후보 비디오들은 생성한 캡션들로 대체된다. 이는 그림 2에서 비디오가 Caption Module로 캡션을 생성하는 부분에 해당한다. i번째 후보 비디오 데이터 캡션을 생성하는 식은 다음과 같다.

$$c_i = Gen(v_i). \tag{2}$$

여기서 Gen는 캡션 생성 모듈, v_i 는 입력 비디오 데이터를 말하며, 결과적으로 c_i 는 i번째 입력 비디오를 텍스트 정보로 표현하는 캡션이다. 이를 통해 생성한 캡션은 입력으로 받는 텍스트와 동일한 형태를 지닌 데이터로 표현되기 때문에 별도의 임베딩 공간 없이도 유사도 계산 커널을 통해 유사성을 계산할 수 있게 된다. 따라서 쿼리로 주어진 캡션과 후보 데이터셋의 각 데이터와의 유사도를 비교하는 과정은 다음과 같다.

$$score_i = \psi(\phi(q), \phi(c_i)). \tag{3}$$

여기서 q는 질의로 주어진 텍스트 쿼리, ϕ는 텍스트 인코더를 의미하고 $\psi(\bullet, \bullet)$는 두 텍스트의 유사도를 비교하기 위한 코사인 유사도 커널을 의미한다. 주어진 쿼리에 대해 이러한 과정을 후보 데이터셋의 모든 후보 비디오에 대해 실행하여 얻은 결과 중 가장 높은 스코어 값을 가지는 후보 비디오를 쿼리에 대한 검색 결과로 판단한다.

기존의 복잡한 시각적 인코딩 모델을 이용하여 임베딩 공간을 학습시켜 추론에 이용한 것과 다르게, 제안 모델은 유사도를 계산할 때 텍스트들을 입력으로 받기 때문에 별도 임베딩 공간에 학습할 필요 없이 추출된 특징을 바로 매칭에 사용할 수 있다. 그러므로 임베딩 공간을 이용하는 모델

에서 지적되어 왔던 이종 간의 데이터 (비디오, 텍스트)가 임베딩 공간에서 표현될 때 분포 차이가 나는 문제 [19]를 완화할 수 있다는 장점을 가진다. 또한 비디오 표현을 위해 시각 인코더 등 많은 모듈을 사용하는 모델 [2, 3] 에 비해 경량화된 모델이기 때문에 연산속도의 향상을 볼 수 있다.

Ⅳ. 실험 결과

4.1 실험 환경

실험을 위해 텍스트-비디오 모델에서 사용하는 대표적인 데이터셋인 MSR-VTT와 VATEX 데이터셋을 사용하였다. MSR-VTT 데이터 세트는 10,000개의 비디오 클립과 이 클립을 설명하는 200,000개의 문장으로 구성되며, 한 클립당 20개의 문장으로 구성되어 있다. VATEX에는 유튜브에서 수집된 10초정도 길이를 가진 34,991개의 동영상 데이터 세트이다. 비디오 한 개당 10개의 영문과 10개의 중국어 문장이 할당되는데 이 실험에서는 영어 문장만 사용한다.

우리는 각각의 학습 데이터 셋을 UniVL을 통해 학습시켰다. 이후 추론 과정에서 테스트 데이터 셋에 대한 캡션들을 만든 뒤 이 캡션들을 비디오의 특성 부분 대신 사용하였다. 실험은 우분투 18.04 환경에서 NVDIA RTX3060 1개를 GPU로 두고 진행하였다. 성능평가의 방법으로 R@K (K = 1, 5, 10), MedR, mAP 까지 총 5개의 순위 기반 측정 지표를 사용했다. R@K는 검색된 상위 K개의 결과 중 하나 이상의 관련 항목이 발견된 테스트 쿼리의 백분율이며, Medr은 검색 결과에서 첫 번째 관련 항목의 중간값이다.

4.2 벤치마크 데이터셋 실험 및 정량적 평가

표 1에서는 MSR-VTT 데이터셋에 대한 실험결과를 보여주고 있다. MSR-VTT를 데이터셋으로 하여 제안모델과 W2VV [20], MEE [1]를 비교 모델로 두어 R@1, R@5, R@10, MedR, mAP를 평가지표로 선정해 실험을 진행하였으며, W2VV와 MEE의 실험결과는 Dual Encoding for Video Retrieval by Text [18]의 실험을 참고하였다.

실험결과를 보면 R@1과 MedR 지표에 대해서는 W2VV, MEE 모델과 다르게 제안 모델이 가장 좋은 성능을 보여주는 것을 확인할 수 있다. 또한 R@5, R@10, mAP지표도 가장 높은 성능을 보이는 MEE모델에 약간 낮은 성능을 보이고 있지만, MEE 모델은 비디오의 특성을 다루기 위해 복잡한 시각 인코딩 과정을 포함하고 있기 때문에 캡션정보로 대체하여 비디오와 텍스트를 한 종류의 데이터로 연산하는 제안 모델이 연산량과 연산 시간에서 강점을 보이며, 다른 모델과 유사하거나 높은 성능을 보인다고 해석할 수 있다.

표 2 에서는 VATEX 데이터셋에 대한 실험결과를 보여주고 있다. VATEX를 데이터셋으로 하여 이전과 동일하게 제안모델과 W2VV [20], VSE++ [21]를 비교 모델로 두어 R@1, R@5, R@10을 평가지표로 선정해 실험을 진행하였으며, W2VV와 VSE++의 실험결과는 Dual Encoding for Video Retrieval by Text [18]의 실험을 참고하였다.

실험결과를 보면 가장 높은 성능을 보이는 VSE++ 모델

표 1. MSR-VTT 데이터셋에 대한 텍스트-비디오 검색 결과
Table 1. Text-video retrieval results on MSR-VTT dataset

Model	R@1	R@5	R@10	MedR	mAP
W2VV [19]	1.1	4.7	8.1	23.6	0.037
MEE [20]	6.8	20.7	31.1	28.0	0.1470
Proposed method	6.9	20.1	29.3	36.0	0.1426

표 2. VATEX 데이터셋에 대한 텍스트-비디오 검색 결과
Table 2.Text-video retrieval results on VATEX dataset

Model	R@1	R@5	R@10
W2VV [19]	14.6	36.3	46.1
VSE++ [21]	31.3	65.8	76.4
Proposed method	31.0	64.0	75.2

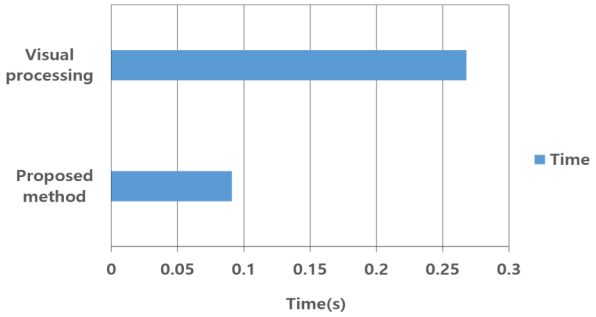


그림 3. 시각정보 처리모델과 제안모델의 텍스트-비디오 검색 시간 비교 결과
Fig. 3. Comparison result of text-video retrieval time between a visual processing model and the proposed model

에 비해 큰 차이를 보이지 않지만, VSE++는 추가적인 샘플링을 통한 학습과정과 시각 인코딩 과정을 거치며 많은 연산이 필요하다. 따라서 캡션 정보를 이용하여 어느 정도 성능을 보장하며 연산량을 줄인 제안 모델이 마찬가지로 큰 강점을 보이는 것을 확인할 수 있다.

그림 3 은 MSR-VTT 데이터셋에 대한 시각 정보 처리 모델과 제안 모델의 텍스트-비디오 검색 시간 즉, 모델의 추론시간 비교 그래프를 보여주고 있다. 시각정보 처리 모델은 제안 모델과 동일한 추출단계를 이용해 비디오에서 시각적인 특징을 추출해 비디오의 특성으로 이용하는 모델이다. 1개의 쿼리 텍스트를 입력해 텍스트의 내용을 가장 잘 담는 비디오를 검색하는데 걸리는 시간까지를 측정하여 비교한 결과를 보여주고 있는데, 제안모델이 0.09초 듀얼 인코딩 모델이 0.26초가 측정되어 제안모델이 시각정보를 처리하는 모델에 비해 약 3분의 1 적은 검색 시간 값을 가진다. 이러한 결과는 임베디드 환경 등의 연산 속도나 연산량이 중요한 분야에서 비디오 검색 문제를 다룰 때 복잡한 비디오 시각처리 인코더 대신 캡션 정보를 사용할 수 있다는 가능성을 시사한다.

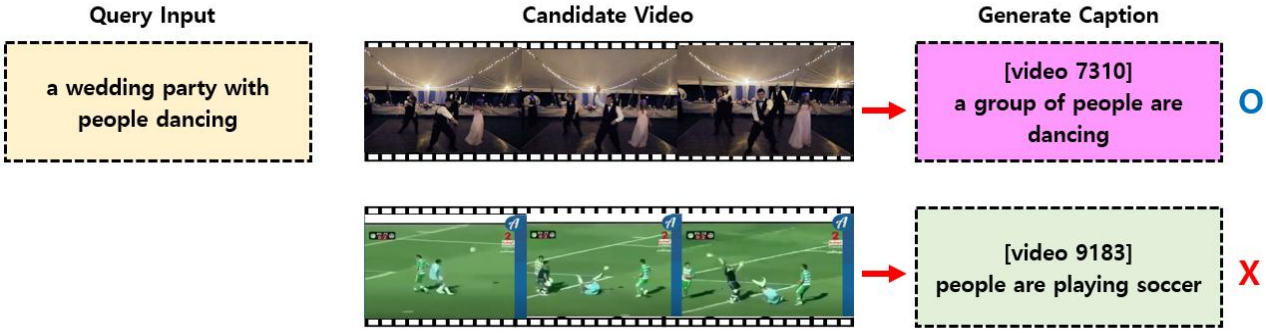


그림 4. MSR-VTT 데이터셋의 추론 결과 예시
Fig. 4. Example of inference result on MSR-VTT dataset

4.3 정성적 평가

그림 4 에서는 MSR-VTT 테스트 데이터셋에 넣어서 나온 모델의 추론결과 예시를 보여주고 있다. 텍스트 쿼리와 후보 비디오 셋 (2990개)를 입력으로 넣었을 때 캡션을 이용한 유사도를 산출한 영상의 결과를 보여주고 있으며, 쿼리로 'a wedding party with people dancing'를 넣은 결과 video7310을 통해 만든 캡션인 'a group of people are dancing'이 가장 높은 score를 얻어 다른 비디오 특징보다 검색 결과에 가깝다고 예측하였고 정답 비디오 클립과 일치하였다. 이러한 결과를 통해 video의 시각적 인코더를 사용하지 않더라도, 캡션 생성 모델을 통하여 비디오를 잘 표현하는 텍스트를 얻을 수만 있다면 텍스트-비디오 검색에서 좋은 특징으로 사용가능함을 볼 수 있다.

V. 결 론

본 논문에서는 비디오를 이용해 만든 캡션 정보를 비디오의 특성으로 활용해 텍스트-비디오 검색 문제를 해결하는 모델을 제안하였다. 기존모델의 경우 비디오의 특성을 처리하기 위해 다양한 처리 방법을 사용한 탓에 두 데이터간 차이에서 오는 격차 문제나 연산량이 복잡하다는 단점을 가지고 있었다면, 제안 모델에서는 비디오의 정보를 담아낼 수 있는 캡션을 생성하고, 비디오의 특성을 캡션으로 정의함으로써 텍스트의 특성과 직접적인 비교를 할 수 있어 공통 임베딩 공간을 구축할 필요가 없다는 장점을 가진다. 또한 많은 양의 연산을 요구하는 비디오의 시각 인코딩 과정을 생략함으로써 기존의 모델 대비 경량화된 장점을 보여주었다.

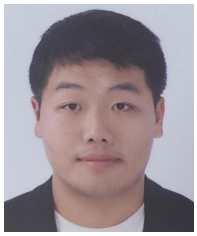
제안 모델을 더 개선하기 위해, 텍스트 특징 추출을 큰 코퍼스를 이용해 학습된 BERT [22] 등의 사전 학습한 모델을 활용한다면 더욱 좋은 성능을 낼 수 있을 것으로 기대된다. 또한 생성 모델을 이용한 캡션이 의미 있는 정보로 활용할 수 있다는 것을 실험을 통해 확인하게 되었고, 이를 비디오의 시각적 정보와 결합하여 더 발전된 검색 모델을 만들어 낼 수 있는 가능성을 보여준다.

References

[1] A. Miech, I. Laptev, J. Sivic, "Learning a Text-video Embedding from Incomplete and Heterogeneous Data," arXiv preprint arXiv:1804.02516, 2018.
[2] N. C. Mithun, J. Li, F. Metze, A. K. Roy-Chowdhury, "Learning Joint Embedding with Multimodal Cues for Cross-modal Video-text Retrieval," in ICMR, pp. 19-27, 2018.
[3] X. Li, C. Xu, G. Yang, Z. Chen, J. Dong, "W2VV++: Fully Deep Learning for Ad-hoc Video Search," in ACM Multimedia, pp. 1786-1794, 2019.
[4] A. Torabi, N. Tandon, L. Sigal, "Learning Language-visual Embedding for Movie Understanding with Natural-language," arXiv preprint arXiv:1609.08124, 2016.
[5] G. Awad, J. Fiscus, D. Joy, M. Michel, A. Smeaton, W. Kraaij, G. Quenot, M. Eskevich, R. Aly, R. Ordelman, G. Jones, B. Huet, M. Larson, "TRECVID 2016: Evaluating Video Search, Video Event Detection, Localization, and Hyperlinking," in TRECVID Workshop, 2016.
[6] X. Yang, J. Dong, Y. Cao, X. Wang, M. Wang, T. S. Chua, "Tree-augmented Cross-modal Encoding for Complex-query Video Retrieval," in SIGIR, pp. 1339-1348, 2020.
[7] A. Miech, D. Zhukov, J. B. Alayrac, M. Tapaswi, I. Laptev, J. Sivic, "Howto100m: Learning a Text-video Embedding by Watching Hundred Million Narrated Video Clips," in ICCV, pp. 2630-2640, 2019.
[8] M. Wray, D. Larlus, G. Csurka, D. Damen, "Fine-grained Action Retrieval Through Multiple Parts-of-speech Embeddings," in ICCV, pp. 450-459, 2019.
[9] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, "Attention is all you need," in NIPS, pp. 5998-6008, 2017.
[10] H. Luo, L. Ji, B. Shi, H. Huang, N. Duan, T. Li, J. Li, T. Bharti, M. Zhou, "Univl: A Unified Video and Language Pre-training Model for Multimodal Understanding and Generation," arXiv preprint arXiv:2002.06353, 2020.

[11] B. Pan, H. Cai, D. A. Huang, K. H. Lee, A. Gaidon, E. Adeli, J. C. Niebles, "Spatio-temporal Graph for Video Captioning with Knowledge Distillation," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10870-10879, 2020.
[12] W. Pei, J. Zhang, X. Wang, L. Ke, X. Shen, Y. W. Tai, "Memory-attended Recurrent Network for Video Captioning," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8347-8356, 2019.
[13] S. Venugopalan, M. Rohrbach, J. Donahue, R. J. Mooney, T. Darrell, K. Saenko, "Sequence to Sequence - Video to Text," in Proc. IEEE Int. Conf. Comput. Vis., pp. 4534-4542, 2015.
[14] L. Gao, Z. Guo, H. Zhang, X. Xu, H. T. Shen, "Video Captioning with Attention-based LSTM and Semantic Consistency," IEEE Trans. Multimedia, Vol. 19, No. 9, pp. 2045-2055, 2017.
[15] J. Song, Z. Guo, L. Gao, W. Liu, D. Zhang, H. T. Shen, "Hierarchical Lstm with Adjusted Temporal Attention for Video Captioning," arXiv preprint arXiv:1706.01231, 2017.
[16] L. Zhou, Y. Zhou, J. J. Corso, R. Socher, C. Xiong, "End-to-end Dense Video Captioning with Masked Transformer," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., pp. 8739-8748, 2018.
[17] L. Huang, W. Wang, J. Chen, X. Wei, "Attention on Attention for Image Captioning," in Proc. IEEE Int. Conf. Comput. Vis., pp. 4634-4643, 2019.
[18] J. Dong, X. Li, C. Xu, X. Yang, G. Yang, X. Wang, M. Wang, "Dual Encoding for Video Retrieval by Text," IEEE Transactions on Pattern Analysis and Machine Intelligence . Vol. 44, No. 8, pp. 4065-4080, 2021.
[19] X. Wang, L. Zhu, Y. Yang, "T2vlad: Global-local Sequence Alignment for Text-video Retrieval," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5079-5088, 2021.
[20] J. Dong, X. Li, C. G. Snoek, "Predicting Visual Features from Text for Image and Video Caption Retrieval," IEEE Transactions on Multimedia, Vol. 20, No. 12, pp. 3377-3388, 2018.
[21] F. Faghri, D. J. Fleet, J. R. Kiros, S. Fidler, "VSE++: Improved Visual-semantic Embeddings," in BMVC, 2018, pp. 1-13.
[22] J. Devlin, M. W. Chang, K. Lee, K. Toutanova, "Bert: Pre-training of Deep Bidirectional Transformers for Language Understanding," arXiv preprint arXiv:1810.04805 (2018).

Dong-hun Lee (이 동 훈)



2017~Computer Science and Engineering from Kyungpook National University (B.S.)

Field of Interests: Video-Text Retrieval, Video Captioning
Email: hy05205@naver.com

Chan Hur (허 찬)

2019~2021 Computer Science and Engineering from Kyungpook National University (M.S.)
2021~Computer Science and Engineering from Kyungpook National University (Ph.D.)

Field of Interests: Video-Text Retrieval, Multimodal Learning
Email: chanhur94@gmail.com

Hyeyoung Park (박 혜 영)

2004~Computer Science and Engineering from Kyungpook National University (Prof.)

Career:
2000~2004 Researchers at Brain Science Institute, RIKEN, Japan
2009~2010 Visiting Professor, Texas A&M University
Field of Interests: Neural Networks, Learning Theory
Email: hypark@knu.ac.kr

Sang-hyo Park (박 상 효)

2011 Computer Engineering from Hanyang University (B.S.)

2017 Computer Science from Hanyang University (Ph.D)

Career:

2017~2018 Postdoctoral position, Korea Electronics Technology Institute (KETI)

2018 Research Fellow, Yonsei University

2019~2020 Postdoctoral position, Ewha Womans University

2020~ Assistant Professor with the School of Computer Science and Engineering, Kyungpook National University

Field of Interesst: VVC, encoding complexity, immersive video, model optimization

Email: s.park@knu.ac.kr