캡션 정보를 이용한 듀얼-인코딩 기반 텍스트-비디오 검색 모델

이동훈†, 허찬†, 박혜영, 박상효

경북대학교

{hypark, s.park}@knu.ac.kr

Dual Encoding Based Text-Video Search Model Using Caption Information

Dong Hun Lee, Chan Hur, Hyeyoung Park, Sang-hyo Park
Kyungpook National University

요 약

본 논문은 비디오-캡셔닝 모델을 이용하여 비디오를 의미적으로 설명하는 텍스트 캡션을 비디오의 특징으로 이용하는 텍스트-비디오 검색 모델을 제안하였다. 제안 모델은 Resnet 인코더와 같은 많은 연산량이 드는 비디오의 시각적 정보 처리 모듈을 텍스트로 대체함으로써 별도의 임베딩 공간을 구축하지 않기 때문에 연산의 경량화 및 연산 시간을 감소시켰다. 제안 모델의 결과 및 효과를 두가지 벤치마크 데이터셋에서 실험을 통해 검증하였다.

Ⅰ. 서 론

텍스트-비디오 검색은 텍스트가 쿼리로 주어질 때 비디오의 내용을 이해하여 이와 의미적으로 일치하는 비디오를 검색하는 연구이다. 이러한 문제를 해결하기 위해 많은 연구들[1,2,3]이 제안되었는데 주로 특징 추출 모델을 이용하여 비디오와 텍스트를 인코딩한 뒤 공통 임베딩 공간상에서 유사도를 비교하는 임베딩 기반 접근법을 사용하고 있다. 또한 비디오의 특성을 인코딩하는 과정에서 object detector, OCR 등의 다양한 정보 추출 모듈을 함께 이용하기도 한다. 하지만 이러한 방법들은 입력에서 비디오와 텍스트, 다른 두 종류의 데이터를 가지고 임베딩 공간을 구성하여 학습을 진행하기 때문에 두 데이터의 형태가 다른 데서 오는 격차 문제를 가지고 있으며 여러가지의 특징추출 모듈을 사용하기 때문에 학습 및 추론 단계에서 많은 양의 연산을 필요로 한다.

우리는 이러한 문제를 해결하기 위해 비디오의 특성을 텍스트로 표현하는 캡셔닝 모듈을 이용하였다. 제안 방법의 장점으로는 비디오의 특성을 캡션으로 정의함으로써 입력으로 넣는 쿼리 텍스트의 특성과 직접적인 비교를 할 수 있어 두가지 다른 데이터의 형태로부터 오는 격차 문제를 완화할 수 있다. 또한 추론 과정에서 가장 많은 연산량이 드는 시각적 인코딩 프로세스를 생략함으로써 모델의 크기 및 연산시간이 줄어들게 된다.

Ⅱ. 본론

2.1 비디오-텍스트 검색 모델

본 논문에서는 비디오-텍스트 검색을 위한 모델로 듀얼-인코딩[4]을 사용하였다. 듀얼-인코딩은 비디오와 텍스트를 입력으로 받아 3 단계의 특징추출과정을 거쳐 얻은 벡터를 공통 임베딩 공간에 매핑하여 유사도를 활용해 탐색하는 모델이다. 우리는 정교한 텍스트 특징 추출을 위해 위의 모델을 텍스트 인코더로 사용하였다.

캡션 데이터 처리를 위한 텍스트 인코딩 과정은 크게 세 단계로 구분이 된다. 우선 특정 문장이 주어지면 이를 원-핫 벡터로 표현한다. 그러고 첫번째 단계(L_1)에서 이 벡터의 평균을 구해 텍스트를 전체적인 정보를 표현하는 글로벌 인코딩 값을 얻는다. 다음, 두번째 단계(L_2)에서는 bi-GRU 를 사용하여 텍스트의 순차적인 정보를 표현하는[1] 벡터를 얻는다. 세번째 단계(L_3)에서는 Bi-GRU 와 1-D CNN 에 기반한 인코딩을 사용하여 지역정보를 구한다. 마지막으로 앞서 계산한 세 단계에서 나온 값들을 결합하여 최종 결과 값들을 얻는다.

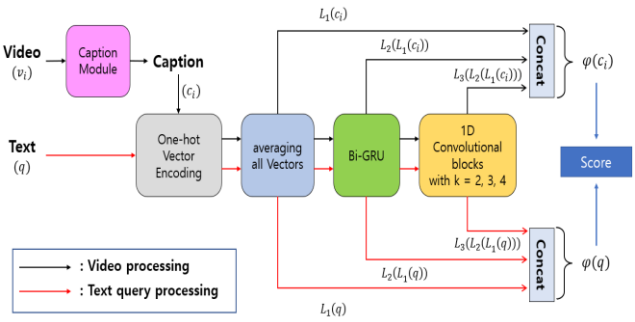

그림 1. 제안하는 텍스트-비디오 검색 모델

2.2 비디오-캡셔닝 모듈

본 논문에서 비디오를 텍스트로 설명하는 인코더-디코더 기반 캡셔닝 모듈 UniVL[5]을 사용하였다. UniVL 을 통해 캡션을 만드는 과정은 다음과 같다. 비디오와 텍스트를 각각 트랜스포머 인코더로 처리를 한 다음, 전처리한 비디오와 텍스트를 결합해 또 다른 트랜스포머를 기반하여 만든 교차 인코더에 넣는다. 마지막으로 인코딩 한 값을 디코더에 넣어 최종적인 캡션을 생성하게 된다. 이때 전처리 과정에서 캡셔닝 모듈의 성능 향상을 위해 데이터셋의 학습 셋을 이용하여 파인-튜닝 과정을 거치게 된다.

†:These authors contributed equally to this work

2.3 캡션정보를 이용한 추론 과정

추론 과정에서 후보 비디오 데이터셋 $D_{test} = \{v_1, v_2 \dots v_i\}$ 은 비디오-캡셔닝 모듈로 생성한 캡션들로 대체된다. 캡션을 생성하는 식은 다음과 같다.

$$c_i = gen(v_i) \quad (1)$$

gen 는 캡션 생성 모듈, v_i 는 입력 비디오 데이터를 말하며, c_i 는 입력한 비디오를 텍스트 정보로 표현하는 캡션이 된다. 생성한 캡션은 입력으로 받는 텍스트와 동일한 형태를 지닌 데이터로 표현되기 때문에 별도의 임베딩 공간 없이도 유사도 계산 커널을 통해 유사성을 계산할 수 있게 된다.

쿼리로 주어진 캡션과 후보 데이터셋의 각 데이터와의 유사도를 비교하는 과정은 다음과 같다.

$$score_i = \psi(\varphi(q), \varphi(c_i)) \quad (2)$$
$$\varphi(t) = [L_1(t), L_2(L_1(t), L_3(L_2(L_1(t)))]$$

여기서 q 는 질의로 주어진 텍스트 쿼리, φ 는 텍스트 인코더를 의미하고 $[\cdot,\cdot]$ 은 concatenation 연산자, L_1, L_2, L_3 은 Dual encoder 모델[4]에서 사용하는 3 단계 텍스트 인코딩 과정, $\psi(\cdot,\cdot)$ 는 유사도를 비교하기 위한 코사인 유사도 커널을 의미한다. 주어진 쿼리에 대해 이러한 과정을 D_{test} 의 모든 후보 데이터에 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 모델에서 지적되어 왔던 이종간의 데이터(비디오,텍스트)가 임베딩 공간에서 표현될 때 분포 차이가 나는 문제[8]를 완화할 수 있다는 장점을 가진다. 또한 비디오 표현을 위해 시각 인코더 등 많은 모듈을 사용하는 모델 [1,2] 에 비해 경량화된 모델이기 때문에 연산속도 향상을 기대할 수 있다.

2.4 실험 결과 및 고찰

우리는 실험을 위해 VATEX 데이터셋을 사용하였다. VATEX 데이터 셋은 유튜브에서 수집된 10 초정도 길이를 가진 34,991 개의 동영상-설명문 쌍으로 구성된 데이터셋이다. 성능평가의 방법으로 R@1, R@5, R@10 까지 총 3 개의 지표를 사용하였다.

Model	R@1	R@5	R@10
W2VV[6]	14.6	36.3	46.1
VSE++[7]	31.3	65.8	76.4
제안 모델	**31.0**	**64.0**	**75.2**

표 1. VATEX 데이터셋에 대한 텍스트-비디오 검색 결과

표 1 에서는 VATEX 데이터셋에 대한 실험결과를 보여주고 있다. 실험결과를 보면 가장 높은 성능을 보이는 VSE++ 모델에 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.

Ⅲ. 결론

본 논문에서는 비디오의 특성으로 비디오를 이용해 만든 캡션 정보를 이용해 텍스트-비디오 검색 문제를 해결하는 모델을 제안하였다. 제안 모델에서는 비디오의 정보를 담아낼 수 있는 캡션을 생성하고, 비디오의 특성을 캡션으로 정의함으로써 텍스트의 특성과 직접적인 비교를 할 수 있어 별도의 임베딩 공간을 구축할 필요가 없다는 장점을 가진다. 또한 많은 양의 연산을 요구하는 비디오의 시각 인코딩 과정을 생략함으로써 기존의 모델 대비 경량화된 장점을 보여주었다. 향후 더 정교한 텍스트 특징추출 모듈을 이용하거나 혹은 대용량 코퍼스로부터 기학습된 텍스트 모델을 이용하는 경우 제안 모델의 성능을 더욱 향상시킬 수 있을 것으로 기대된다.

ACKNOWLEDGMENT

이 논문은 2020 년도 정부(교육부)의 재원으로 한국연구재단의 지원(No. 2020R1I1A3072227, 기초연구사업임)을 받아 수행된 연구임.

참 고 문 헌

[1] Multi-modal Transformer for Video Retrieval (Gabeur, Valentin, et al. ECCV, 2020)

[2] Use What You Have_ Video Retrieval Using (Liu, Yang, et al. , BMVC, 2019)

[3] Attention Mechanisms, Signal Encodings and Fusion Strategies (Galanopoulos, Damianos, and Vasileios Mezaris. ICMR, 2020)

[4] Dual Encoding for Video Retrieval by Text (Dong, Jianfeng, et al. TPAMI, 2021)

[5] UniVL: A Unified Video and Language Pre-Training Model for Multimodal Understanding and Generation (Luo, Huaishao, et al. 2020)

[6] Predicting visual features from text for image and video caption retrieval (Dong, Jianfeng, Xirong Li, and Cees GM Snoek. IEEE, 2018)

[7] VSE++: Improved visual-semantic embeddings with Hard Negatives (Faghri, Fartash, et al. BMVC, 2018)

[8] T2VLAD_Global-Local Sequence Alignment for Text-Video Retrieval (Wang, Xiaohan, Linchao Zhu, and Yi Yang. CVPR, 2021)