Video retrieval, captioning 기반 동영상 필터링 모델

김창민, 김정욱
경희대학교 컴퓨터공학과
kcmin19@khu.ac.kr

Video retrieval, captioning-based video filtering model

Changmin Kim, Jeongwook Kim
Department of Computer Science and Engineering, KyungHee University

요 약

본 연구에서는 수집된 비디오 데이터의 설명문을 생성하고, 이를 비디오의 text feature로 활용하여 retrieval을 진행한다. 이때, text feature와 비디오 간의 유사도를 이용하여 관련 없는 비디오를 필터링하는 모델을 구현한다. MSR-VTT 데이터셋에 대해 이 방법을 적용한 후 군집화를 통해 비디오를 필터링한 결과, 유사한 비디오는 captioning 모델에 의해 비슷하거나 동일한 설명문을 생성하기 때문에 같은 군집에 속하는 경향이 있고, 영상에 대한 명확한 설명문을 생성하지 못한 경우 비슷한 동영상임에도 다른 군집으로 분류된다.

1. 서 론

빅데이터가 발전해감에 따라 비디오, 오디오, 텍스트, 이미지 등 다양한 데이터가 연구에 활용되고 있다. 이 중 비디오 분야의 연구를 위한 데이터는 직접 caption을 생성해 검증된 데이터를 활용하거나, 범용적으로 사용하기 위해 직접 여러 검색엔진으로부터 확보한다. 그러나 검색엔진은 주로 사용자의 만족도를 최대화하기 위한 목적으로 개발되었기 때문에 관련성이 떨어지거나 현재 유행에 따라 왜곡된 결과가 노출되기도 한다.

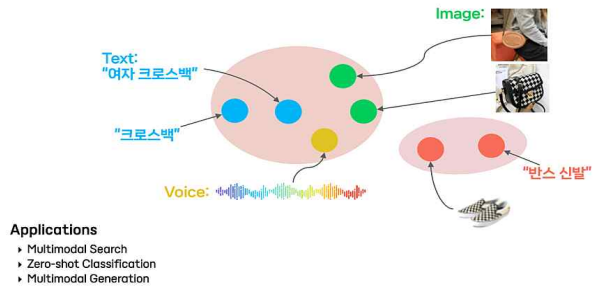
따라서 동영상 데이터 수집 단계에는 labeling이나 전처리 이외에도 관련 없는 데이터를 직접 검수해서 제거하는 과정을 거쳐야 할 필요가 있다. 본 연구에서는 수집된 영상의 captioning 결과를 비디오의 text feature로 처리하여 retrieval을 수행했을 때 검색 결과가 상호 연관성이 있는지 파악하는 모델을 구현하고자 한다.

2. 관련 연구

2.1 Multi-modal

Modality는 소통 과정에서의 다양한 전달 방식을 의미한다. 전통적인 딥러닝도 이를 기반으로 이미지, 텍스트, 음성 등 단일 Modality에 기반한 모델이 주로 개발되고 있다. Multi-modal은 여러 다른 Modality를 함께 활용하여 서로의 관련성을 학습하고 여러 형태로 표현하는 기술이다. Multi-modal에 기반한 모델은 입출력에 있어 하나의 Modality를 학습한 모델에서보다 더 유연한 처리가 가능하기 때문에 기존보다 다양한 작업을 수행할 수 있다.

이미지, 텍스트 등이 representation 영역에서 가까운 위치에 있으면 같거나 유사한 내용을 담고 있다는 가정을 하고 하나의 모달리티로 다른 모달리티를 검색하는 retrieval이나 이전에 본 적 없는 데이터의 class를 예측하는 zero-shot classfication, 시각적 미디어 컨텐츠에 대한 텍스트 설명을 생성하는 captioning 등 현재는 Multi-modal을 통해 사람의 감각과 맞닿은 소통이 가능한 AI가 개발되고 있다.



[그림1 Multi-modal representation[1]]

2.2 군집화

군집화는 비지도 학습에 속하는 알고리즘으로 유사한 성질을 가지는 개체를 그룹으로 묶는 기법이다. 개체 간의 거리, 유사도 등을 통해 개체를 하나의 그룹으로 묶으며, 거리를 기반으로 한 군집화 기법에는 군집 내에 속한 개체의 중심을 나타내는 군집 중심을 가지고 있다.

군집화는 별도로 label을 할당받지 않은 데이터에 유사도를 기반으로 label을 선정하거나, 데이터 중 유사하지 않은 데이터들을 색출해내는데 사용된다. 하지만 군집화는 여러 문제점을 같이 지니고 있다. K-means와 같은 기법은 초기 중심점에 따라 군집화 실행 시마다 다른 결과를 도출하기도 하고, 이상치를 모두 군집에 포함시키는 군집화 알고리즘은 노이즈가 크게 작용할 수 있다. 또한 차원이 많고 범위가 넓은 경우에, 유사도를 파악하기 위한 각 특성의 영향력을 선정하는데도 어려움이 있으며, 모델 동작에 큰 시간이 걸리는 문제점이 있다.

본 논문에서는 프로젝트의 진행에 DBSCAN[2] clustering을 사용한다. DBSCAN은 밀도를 기반으로 한 공간 군집화 알고리즘으로, 군집을 이룰 수 있는 최소거리와 군집을 이루기 위한 최소 노드 수만을 하이퍼 파라미터로 받아 군집화 작업을 수행한다. 군집을 구성하며, 군집 안에 포함될 수 있는 거리에 위치하지 않은 노드는 이상치로 분류되며 특정 군집에 속하지 않는다.

해당 알고리즘은 밀도가 다른 군집을 식별하거나, 다차원의 데이터를 군집화할 때 어려움을 겪을 수 있으며, 연산 복잡성으로 인한 처리 시간이 길어질 수 있다. 이 문제는 동영상에 대한 captioning을 진행하고, 그 caption 값을 기준으로 동영상을 retrieval 했을 시 나타나는 검색 순위, 즉 문장과 비디오의 유사도를 이용하여 군집화를 진행해서 밀도가 고른 낮은 차원의 군집화를 수행함을 통해 문제를 해결할 수 있으므로, 프로젝트에 적합한 기법이라 볼 수 있다.

3. 프로젝트 내용

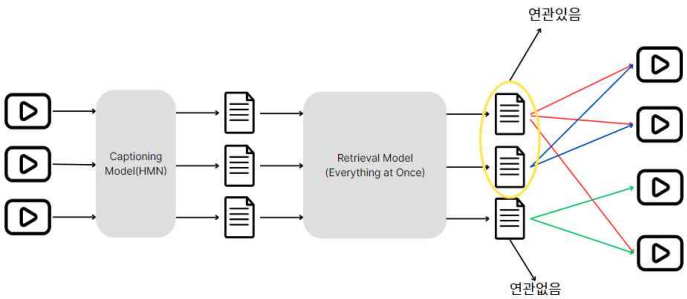
3.1 문제정의

검색에 기반한 동영상 데이터 수집 시 데이터의 제목이나 설명에 의존한 데이터 수집이 이루어지기 때문에 수집된 일부 동영상이 검색어와 다른 내용의 동영상일 수 있다. 이 경우 해당 동영상을 데이터에서 직접 제거하거나 검수하는 작업을 필요로 한다. 본 연구에서는 검색을 통해 사용자가 동영상 데이터를 수집할 때, 일차적으로 필터링하기 위한 모델을 만들고자 한다.

3.2 프로젝트 시나리오

3.2.1 프로젝트 요약

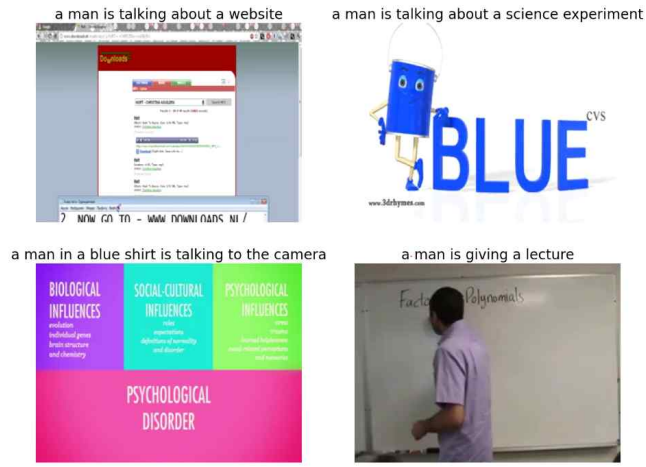
그림2는 해당 프로젝트의 시나리오를 나타낸다. 데이터는 captioning, retrieval 각 모델에 사용된 MSR-VTT[3] 비디오 설명문 데이터를 활용한다. 비디오로부터 각 모델에 맞게 특성을 추출한 후 영상별 captioning을 수행, 생성된 설명문을 각 영상의 text feature로 나타낸 뒤 비디오 내에서 retrieval을 수행한다. 이때, 비디오가 검색결과를 도출하기 위한 각 캡션과 비디오와의 유사도를 추출한다. 본 연구에서는 유사한 동영상으로부터 얻은 영상 설명문은 비슷한 검색 결과를 가져올 것으로 예상하여, 설명문과 비디오의 유사도를 이용한 상관계수를 측정한다. 구한 유사도를 바탕으로 DBSCAN 군집화를 수행하여 결과를 확인한다.



[그림 2 프로젝트 시나리오]

3.2.2 Captioning

Captioning 모델은 오픈소스를 제공하는 HMN[4] 모델을 사용한다. HMN 모델은 비디오에 나타나는 객체를 탐지한 후, 텍스트 설명을 기반으로 각 객체에 맞는 텍스트가 어느 부분인지 학습하고 보다 큰 범위의 장면, 시간적인 구조가 텍스트의 어느 부분과 연관되어있는지를 학습하는 모델이다. 해당 논문에서 제공하는 코드, 추출된 특성을 사용하여 동영상의 설명문을 생성한다. 그 결과 프로젝트에 사용될 총 968개의 비디오에 대해 설명문을 생성했다.



[그림 3 생성된 설명문]

3.2.3 Retrieval

Retrieval 모델은 Everything at Once[5] 논문에서 제공하는 퓨전 트랜스포머 모델을 사용한다. 해당 모델은 비디오나 다른 미디어 데이터로부터 modality 성분을 추출하고 그를 결합한 임베딩 공간을 형성해 modality에 무관한 검색을 수행할 수 있는 retrieval 모델이다. 해당 모델에서는 비디오에 할당된 설명문을 비디오의 text 성분으로 하여 검색을 수행하고, 다른 비디오와의 유사도를 측정한다. 이때, 특성을 추출한 동영상에 할당된 설명문은 앞서 HMN 모델에서 생성한 설명문을 사용하고, 동영상 필터링을 위해 t2v 유사도를 사용한다. 사용되는 유사도는 코사인 유사도이다. 그림 4에서 특정 행렬에 대한 값은 행 비디오의 설명문에 대한 열 비디오의 유사도를 나타낸다. 설명문의 내용은 모델로부터 생성한 값으로 비디오의 내용과 완전히 일치하지는 않고, retrieval 모델의 성능도 정확하지 않기 때문에, 검색 결과가 자기 자신에 대해 가장 높은값을 가지지는 않는다.

	video7020	video7021	video7024	video7025	video7026	video7027	video7028	video7029	video7034	video7035	...
video7020	0.001743	0.023562	0.089442	0.064106	-0.042263	0.136810	0.046964	0.169989	0.088067	0.140991	...
video7021	-0.014428	0.382332	0.168053	0.257322	0.005939	0.168425	0.214407	0.232058	0.203516	0.223945	...
video7024	0.062864	0.148763	0.185027	0.063717	-0.039862	0.176594	0.082787	0.271756	0.155573	0.209092	...
video7025	-0.006881	0.192642	0.120589	0.293928	0.083577	0.151418	0.234543	0.223440	0.133456	0.202993	...
video7026	-0.093191	0.111537	0.212854	0.130998	0.187697	0.117603	0.118072	0.153686	0.125160	0.129372	...
...	...	...	...	...	...	...	...	...	...	...	...

[그림 4 설명문과 비디오의 유사도]

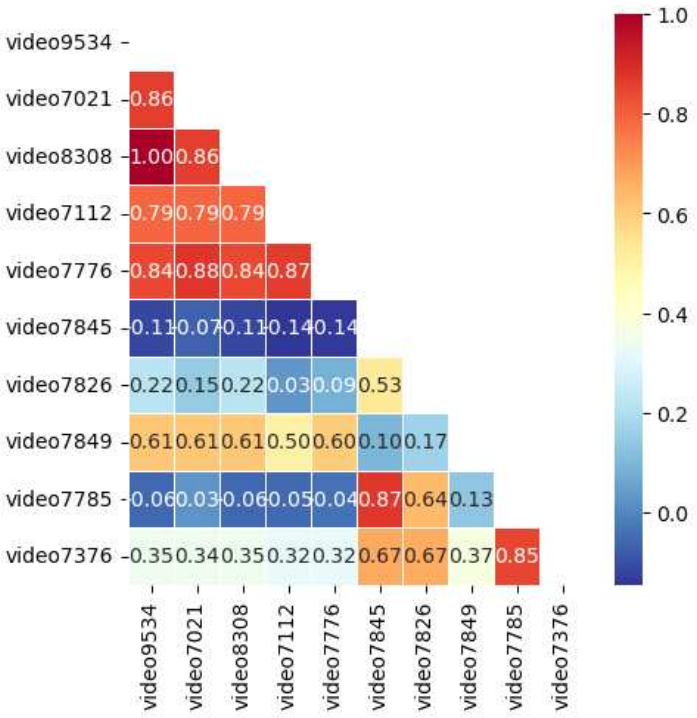
4. 결과 분석

4.1 유사도 분포

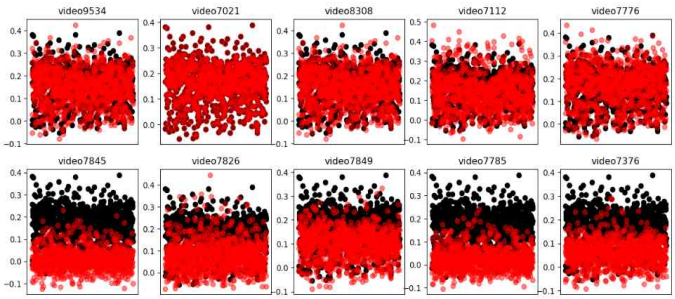
프로젝트 진행에 사용된 데이터는 968개로, 본 프로젝트에서는 retrieval 유사도와 상관계수에 대한 설명을 위한 예시로 “video7021”을 사용한다. 해당 비디오의 기존 caption은 “baseball player hits ball”이고, 새로 생성된 설명문은 “people are playing baseball”이다. 그림 5는 해당 비디오와 가장 높은 유사도를 보이는 3가지 영상(위 3장), 가장 낮은 유사도를 보이는 3가지 영상(아래 3장)의 첫 프레임을 나타내고 있다. 즉, retrieval 모델이 판단하는 people are playing baseball과 가장 유사한 동영상, 가장 유사하지 않은 동영상을 나타낸다. 유사도가 높은 동영상의 경우 야구장의 모습이나 야구선수가 관측되고 있다. 이때 유사도가 높은 video9534, video8308의 경우에도 스스로 생성한 caption을 가지고 다른 비디오에 대해 video7021과 비슷한 유사도를 나타낼 것으로 추정했다.

[그림 5 video7021과 유사한(위) 동영상, 차이가 큰(아래) 동영상]

이를 바탕으로 Video7021과 가장 유사도가 높은 5개, 가장 낮은 5개 동영상의 설명문과의 유사도를 산점도로 나타내었다. 그림 7에서, video7021의 설명문과 다른 968개의 비디오들 간의 유사도는 검은색, 각 비디오의 설명문과 다른 비디오들 간의 유사도는 빨간색으로 표시되었다. 유사도가 높은 그룹의 동영상은 두 분포가 겹치는 경향을 보이며, 유사도가 낮은 그룹은 분포에 차이를 보인다. 이는 비슷한 동영상이 생성한 설명문이 높은 유사도를 가지며, 해당 동영상들의 유사도 분포도 유사함을 시사한다. 다시 말하면, 유사도를 상관계수로 나타내었을 때, 비슷한 동영상은 높은 수치를 보인다고 할 수 있다. 그림 6은 각 해당 10개 비디오의 피어슨 상관계수를 나타낸다.



[그림 6 video7021과 연관된 10개 비디오의 상관계수]



[그림 7 video7021과 연관된 10개 비디오의 유사도 산점도]

4.2 필터링 모델 성능

앞서 구한 상관계수를 이용해 각 비디오를 군집화했다. 군집화 알고리즘은 현재 프로젝트의 목적, 데이터의 구성을 기준으로 선정하였다. 프로젝트는 불필요한 비디오 데이터의 제거, 즉 이상치를 필터링 하는 기법을 요구하고 있으며, 데이터는 968개 행, 968개 열로 구성된 각 동영상 유사도 간의 상관계수이므로, 밀도가 비교적 균일한 데이터로 판단하여 DBSCAN Clustering을 선정했다. 군집화 모델의 하이퍼 패러미터로 군집당 최소 샘플 수를 3으로 하고, 군집을 이루기 위한 최소 거리를 2.5로 했을 때, 총 30개의 군집, 79개의 필터링 된 데이터를 추출했다. 이는 MSR-VTT 데이터셋이 다수의 카테고리로부터 추출된 영상이고, 패러미터의 수치를 높이는 경우 다소 이상치가 포함될 수는 있으나, 필터링 된 데이터를 포함시킬 수 있을 것으로 보인다.

그림 8은 DBSCAN이 분류한 하나의 군집이다. 종이접기와 관련된 동영상으로 분류가 수행되어 모델의 동작이 잘 이루어진 모습을 보인다. 해당 동영상의 설명문을 확인해보면 “a person is folding paper” 으로 동일한 설명문을 가진 비디오가 존재한다. 이는 caption 모델이 비디오에 대해 간단한 설명문을 생성했기 때문이며, 완전히 동일한 설명문으로 유사도를 구하였기 때문에, 같은 군집에 속했을 것으로 예상된다. 하지만 “a person is making a piece of paper” 와 같이 유사한 설명문을 생성하여 같은 군집에 속한 비디오 또한 존재한다.

[그림 8 처리 결과1]

군집화가 잘 이루어지지 않은 경우도 존재하는데, 이는 captioning 시 모델이 비디오에 대한 설명문을 상세하게 작성하지 못하고 큰 분류의 feature만 두고 설명문을 생성해서 서로 다른 여러 동영상이 같은 설명문을 가지게 되어 문제가 발생한 것으로 추정된다. 그림 9는 모델이 생성한 군집 중 하나로, 영상과 일치하는 caption이 생성되지 않았거나, 내용은 다르나 유사한 특성을 보유하여 하나의 군집으로 분류되었다.

[그림 9 처리 결과2]

5. 결론

본 연구는 비디오 데이터를 수집하는 과정에서의 이상 데이터를 파악하고 직접적인 검수가 없어도 이상치를 분류해낼 수 있는 모델을 구현하고자 한다. 비디오에 포함된 Multimodal에 대해 Captioning과 Retrieval 기법을 적용하여 광범위한 feature와 차원을 유사도로 압축한 결과를 바탕으로 군집화를 수행하였다.

현재 모델에서는 다른 동영상에서 동일한 설명문을 보이는 등 retrieval 모델의 성능 보다는 captioning 모델의 성능이 더 중요한 영향을 미치고 있으며, 설명문 생성 능력이 증가하면 더 섬세한 이상치 분류 능력을 보일 것으로 예상된다.

일반적인 비디오 분류의 경우 학습을 위해 동영상의 label이 필요하며, 동영상의 여러 특성을 추출하고 조합하기 위한 복잡한 연산과정이 필요하기 때문에 새로 수집하는 데이터에 적용하기 어렵다. 하지만 해당 모델의 경우 captioning과 retrieval 모델의 사용으로 zero-shot 분류가 가능하여 비교적 장점을 보이고, multimodal 학습의 고도화에 따라 성능을 끌어올릴 수 있다.

6. 참고문헌

[1] 전동현 “멀티모달 AI가 바꿀 미래”, https://channeltech.naver.com/contentDetail/25
[2] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu, “A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise”, AAAI, 1996
[3] Jun Xu , Tao Mei , Ting Yao and Yong Rui “MSR-VTT: A Large Video Description Dataset for Bridging Video and Language”, CVPR, 2016
[4] Hanhua Ye, Guorong Li, Yuankai Qi, Shuhui Wang, Qingming Huang, Ming-Hsuan Yang “Hierarchical Modular Network for Video Captioning”, CVPR, 2022
[5] Nina Shvetsova “Everything at Once - Multi-modal Fusion Transformer for Video Retrieval”, CVPR, 2022