# CLIP 기반 멀티모달 지원 비디오 검색 기술에 관한 연구

박원주, 이남경
한국전자통신연구원 미디어연구본부 미디어지능화연구실

[wjpark, nklee]@etri.re.kr

# A study on Multimodal Video Retrieval using the CLIP

Wonjoo PARK, Namkyung Lee
Media Intellectualization Research Section

Media Research Division, ETRI

요 약

콘텐츠 기반 비디오 검색은 주어진 사용자 질의와 유사하거나 중복되는 콘텐츠를 비디오 아카이브에서 검색하는 것을 목표로 한다. 이 기술은 최근 비디오 콘텐츠의 생산 시장의 확장으로 실질적인 요구가 커지고 있다. 본 논문은 CLIP 을 활용하여 텍스트, 이미지, 동영상을 활용한 멀티모달 사용자 질의를 지원하는 비디오 검색 기술 및 시스템에 대한 연구 결과를 제안한다. 특히 한국어로 장면 설명이 태깅된 텍스트와 비디오 쌍을 학습하여 크로스 모달 간 데이터를 표현하는 학습 시스템의 경험적 연구 결과를 제시하고, 검색 속도를 향상시키기 위하여 유사 근접 해법 기술을 적용한 멀티모달 지원 비디오 검색 시스템을 소개한다.

## Ⅰ. 서 론

방송 기술의 발전과 보급, 동영상 크리에이터 시장의 성장으로 방송국뿐만 아니라 OTT, Youtube 등 다양한 플랫폼을 통하여 미디어 콘텐츠가 제작 및 배포되고 있다. 특히 전통적인 방송 분야에서는 콘텐츠의 아카이브화 및 메타데이터 구축을 통하여 자사의 기존 영상을 활용하고 검색하려는 노력을 지속해오고 있다. 더 나아가 K 엔터테인먼트의 세계적 위상이 높아지고, K 콘텐츠의 수요가 글로벌로 확장되는 추세에 따라 이미 방영된 콘텐츠를 재활용하고자 하는 목적은 더욱 뚜렷하다. 최근 제작하는 콘텐츠는 이러한 수요에 따라 제작 및 배포시점에 메타데이터를 구축하는 시스템을 도입하여 가급적 풍부한 정보를 함께 아카이브화하고 있다. 그러나, 과거의 콘텐츠들은 이러한 수요없이 제작 및 배포된 환경으로 이를 검색하고 재활용하는 것은 인적자원에 매우 의존적일 수밖에 없다. 따라서 본 논문은 아카이브화된 콘텐츠를 다양한 사용자 질의 방식을 지원하며 검색할 수 있는 기술 및 시스템을 제안한다. 또한 자연어 처리 분야에서 대용량 코퍼스를 사전 학습하여 다양한 과업에 적용하여 우수한 성능을 보일 수 있음을 보여준 BERT 등 모델이 공개된 후, 비디오 처리 분야도 이와 유사한 연구가 진행되고 있다[1]. 특히, 미디어 콘텐츠가 비디오 포맷이고, 사용자 쿼리가 텍스트 포맷인 크로스 모달인 경우를 고려하여, 본 논문은 최근 이미지와 텍스트를 쌍으로 학습하여 텍스트로 이미지를 검색하는 분야에서 좋은 성능을 보인 CLIP 을 비디오로 확장하여 비디오 검색 분야에서 좋은 성능을 보인 CLIP4clip 을 활용하여 비디오 검색 기술에 적용하고자 한다[2][3].
마지막으로 크로스 모달 학습 모델로 표현된 비디오를 DB 에 저장하고, 텍스트, 이미지, 비디오 등 3 종의 멀티모달 인터페이스를 지원하는 시스템을 소개한다. 특히 검색 속도를 향상시키면서, 코사인 유사 수준에 따라 의미적으로 비디오를 검색하기 위하여 트리형 근사 해법 기술을 적용한 시스템 설계 및 개발 결과를 살펴본다.

## Ⅱ. CLIP 기반 멀티모달 지원 비디오 검색 시스템

CLIP(Contrastive Language-Image Pre-training)은은 기존의 이미지 분류 과업에서 대용량의 레이블링 데이터가 필요했던 한계를 극복하고, NLP 분야에서 레이블이 없는 대용량의 코퍼스로 사전학습을 수행하여 자연어 처리의 다양한 분야에서 우수한 결과를 도출한 사례를 적용하여 약 4 억개의 이미지와 텍스트 데이터쌍을 구축하고, 크로스 모달 데이터를 각각 인코딩한 후, 쌍으로 이루어진 두 벡터들의 연결관계를 학습한 사전학습 모델이다. CLIP 모델을 활용한 텍스트 기반 이미지 검색 기술은 기존의 검색 성능을 크게 향상시켰다. 따라서 이를 비디오로 확장하여 CLIP 을 기초로 텍스트 기반 비디오 검색 기술을 적용하는 사례가 출현하였다. CLIP4clip 은 CLIP 사전학습 모델을 기초로, 비디오를 사용자 파라미터에 따라 프레임을 추출하고, 추출된 프레임 이미지를 2D, 또는 3D 패치하여 텍스트와 쌍으로 학습한다. CLIP4clip 은 다양한 공개 데이터로 실험하여 텍스트 기반 비디오 검색 기술에서 우수한 성능을 보였다.
본 논문은 한국어 질의 기반 비디오 검색을 위하여 한국영화 250 편을 수집하고, 20 초단위의 클립으로 분할하였다. 전체 클립의 수는 79,204 개이고, 시간은 약 90 만초이다. 영화는 드라마, 액션, 코미디 등 다양한 장르를 포함하고 있으며, 분할된 클립에는 50 자 이하의 한국어로 장면을 서술하는 문장을 태깅하여 활용하였다[4]. CLIP4clip 은 텍스트와 비디오 포맷의 데이터를 각각 인코딩하여 크로스 모달간 유사도를 학습하기 위하여 각 포맷에 적용가능한 인코딩 모델이 필요하다. 비디오의 경우 언어의 의존도가 없으므로 CLIP 에서 ResNet 보다 좋은 성능을 보인 ViT-B/32 를 적용하고, 텍스트의 경우 한국어가 지원 가능한 인코더를 위하여 multilingual 모델을 활용한다. 궁긍적으로 CLIP 과 유사하게 텍스트와 비디오를 공통의 임베딩 공간에 미디어 시맨틱을 인코딩하여 표현하고, 비디오와 텍스트 쌍의 인코딩 벡터 간 코사인 유사도가 최대화되도록 연결관계를 학습하는 미세 조정 학습을 수행한다. 이렇게

학습된 모델을 활용하여 검색 데이터를 대상으로 인코딩하여 DB 에 저장한다. DB 에 저장하고 색인 시, 검색 속도를 향상시키기 위하여 유사 근접 해법을 적용한다. NN(Nearest Neighbor)는 최근접 이웃 알고리즘으로 공간에서 가장 가까운 벡터를 찾는 기법이다. 최근 추천 시스템에서 질의 벡터에 대해서 유사한 벡터를 추천해줄 때 주로 사용되는데, 최근접 이웃을 찾는 것은 확장성을 고려할 때 시간이 많이 소요되는 알고리즘이다. 이럴 때 유사 근접 벡터를 탐색하는 기법으로 ANNOY(Approximate Nearest Neighbors Oh Yeah) 를 적용한다[5]. ANNOY 는 spotify 에서 음악 추천을 위하여 주어진 질의점과 가까운 임베딩 지점을 검색해주는 라이브러리이다.

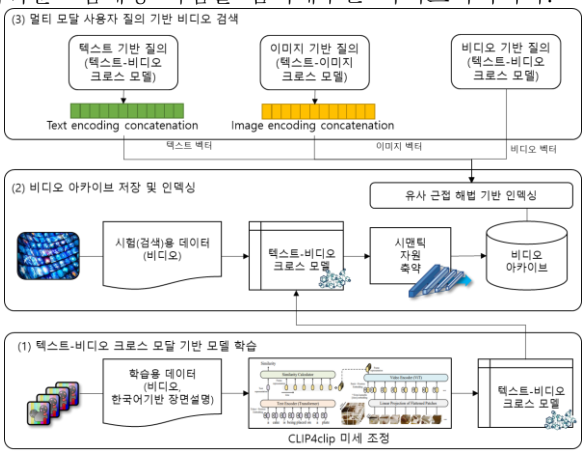


그림 1. 멀티모달 지원 비디오 검색 시스템

그림 1 은 CLIP4clip 기반으로 비디오와 텍스트 쌍을 공통된 임베딩 공간에 표현하는 모델 학습과 이 모델을 활용하여 유사 근접 해법을 통하여 DB 에 적용하고 인덱싱하여 텍스트 기반 사용자 질의를 지원하여 비디오를 탐색하는 검색 시스템이다. 특히 시스템은 텍스트-비디오 검색뿐만 아니라, CLIP 모델을 활용하여 이미지 질의 기반으로 이미지가 포함된 비디오나 유사 비디오을 검색하는 이미지-비디오 검색을 지원한다. 이미지를 임베딩하고 유사도에 따라 검색하기 위하여 이미지 인코딩 벡터를 순차적으로 반복 연결하여 비디오 벡터의 길이에 맞추어서 질의한다. 또한 시스템은 비디오 특징이 유사한 영상 검색을 위하여 비디오-비디오 검색을 제공한다. 표 1 은 CLIP4clip 크로스 모델 미세 조정 학습 실험 환경이다. 비디오 프레임 수는 학습 대상 비디오에서 추출한 이미지 프레임의 수이고, 비디오 추출 방법은 비디오에서 이미지 프레임을 추출할 때 비디오에서 균등하게 이미지를 추출방법이다. 텍스트 최대 단어의 수는 언어 모델에 임베딩할 때 최대 텍스트의 길이이다. 한국어 장면 설명의 경우 비디오 검색 의도에 맞게 가능하면 시간과 장소와 행위와 행위 주체 등을 포함하고 있다.

표 1. 학습 실험 환경

| 구분 | 환경 |
|---|---|
| 비디오 프레임 수 | 12 개 |
| 비디오 추출 방법 | 균등(uniform)추출 방법 |
| 텍스트의 최대 단어 수 | 32 개 |
| 한국어 장면 설명의 예 | 방송국에서 이야기하는 남자 |

표 2. 모델 실험 결과

| 구분 | R@1 | R@5 | R@10 | Median Rank |
|---|---|---|---|---|
| 텍스트-비디오 | 29.2 | 59.6 | 69.9 | 4.0 |

표 2 의 R@1, R@5, R@10 은 사용자 질의에 대하여 상위 1, 5, 10 결과에서의 재현률이고, Median Rank 은 검색 결과에서 정답의 중앙값을 의미한다. 같은 모델을 활용하여 영문으로 학습하고 검색한 결과로 각각 44.5, 71.4, 81.6, 2.0 의 성능을 보이고 있다. 본 시스템의 학습 결과를 분석하면 텍스트 질의기반 비디오를 검색하는 경우 상위 10 개안에 70%에 가까운 재현율을 보이고 있고, 검색 결과의 중앙값도 4.0 을 보이고 있는데, multilingual 모델을 활용한 것을 고려하였을 때, 비교적 우수한 성능 달성하였다. 그림 2 는 위의 모델을 유사 근접 해법을 적용하여 DB 에 저장하고, 텍스트 포맷으로 비디오를 검색한 결과의 예이다.

“질의어 : 스튜디오에서 전화를 받는 남자”

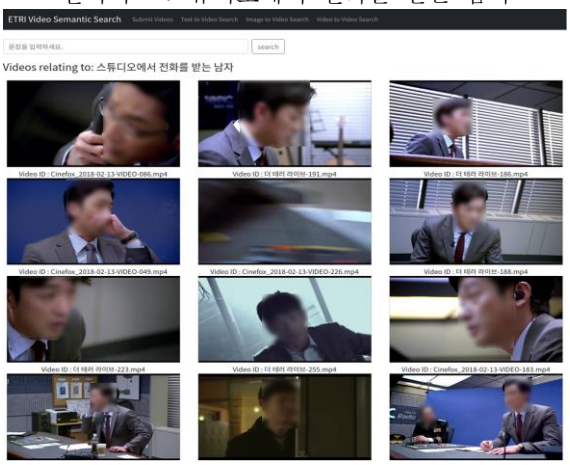


그림 2. 텍스트 기반 비디오 검색 결과의 예

## Ⅲ. 결 론

본 논문은 텍스트, 이미지, 비디오 포맷으로 비디오를 검색하는 기술 및 시스템 결과를 제안하였다. 한국어로 장면 설명이 태깅된 비디오쌍을 CLIP4clip 을 미세 조정 학습하여 모델링하고, 유사 근접 해법 기반으로 DB 에 저장 후 인덱싱하여 사용자 질의에 적절한 검색 결과를 보여주었다. 향후 본 기술은 콘텐츠의 확장 및 텍스트-동영상 크로스 모달 학습 방법 개선, 확장성 제공이 가능한 검색 기법을 적용한 시스템을 개발할 계획이다.

## ACKNOWLEDGMENT

이 논문은 2022 년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (No. 2021-0-00852, 지능적 미디어 속성 생성 및 활용 기술 개발).

## 참 고 문 헌

[1] Dosovitskiy, Alexey, et al. "An image is worth 16x16 words: Transformers for image recognition at scale." arXiv preprint arXiv:2010.11929 (2020).

[2] Radford, Alec, et al. "Learning transferable visual models from natural language supervision." International Conference on Machine Learning. PMLR, 2021.

[3] Luo, Huaishao, et al. "Clip4clip: An empirical study of clip for end to end video clip retrieval." arXiv preprint arXiv:2104.08860 (2021).

[4] K. Hahm, C. Kwak, and S.Kim. "Learning a Video-Text Joint Embedding using Korean Tagged Movie Clips." 2020 International Conference on Information and Communication Technology Convergence (ICTC). IEEE, 2020.

[5] https://github.com/spotify/annoy