TAVE 16기 심화프로젝트 무비탐정팀(박지혜, 박주연, 남경민)

1. 프로젝트 개요 및 목적

유튜브나 넷플릭스와 같은 영상 스트리밍 서비스는 방대한 영상 콘텐츠를 제공하지만, 영상의 길이가 길어질수록 사용자가 특정 장면을 다시 찾는 과정은 매우 비효율적이다. 현재 사용자는 원하는 장면을 찾기 위해 재생 바를 앞뒤로 반복 이동하며 시간대별 장면을 직접 확인하거나, 다른 사용자가 댓글로 정리해 둔 타임라인을 참고하는 방식에 의존하고 있다. 이러한 방식은 단순한 시간 정보에 기반한 탐색에 그치며, "누가 등장하는 장면", "긴장감이 고조되는 순간", "특정 사건이 발생하는 장면"과 같이 영상의 의미적 내용을 기반으로 한 탐색에는 근본적인 한계가 있다. 특히 영화와 같이 서사 구조와 인물 관계가 중요한 영상 콘텐츠에서는 의미 중심의 질의로 원하는 장면을 찾는 것이 사실상 불가능하다.

본 프로젝트는 이러한 실생활에서의 불편함에서 출발하여, 영상의 의미를 이해하고 이를 기반으로 장면을 검색할 수 있는 시스템을 구축하는 것을 목표로 한다. 영화 장면을 단순한 시간 단위가 아닌 이미지·대사·등장 인물 정보를 포함한 의미 단위로 표현하고, 사용자가 자연어로 입력한 질의에 대해 의미적으로 가장 적합한 장면을 검색하는 의미 기반 영상 장면 검색 시스템을 제안한다. 이를 통해 사용자는 영상의 길이나 정확한 시간 정보를 알지 못하더라도 기억 속의 장면이나 상황을 자연어로 표현하여 원하는 장면을 직관적으로 탐색할 수 있으며, 기존 영상 탐색 방식의 한계를 효과적으로 보완할 수 있다.

2. 데이터 소개(데이터셋, 수집 방법 및 출처)

본 프로젝트에서 사용한 영상 데이터는 영화 기생충의 블루레이 디스크를 직접 구매하여, 개인 연구 목적으로 로컬 환경에 저장한 영상 파일을 기반으로 구성하였다. 저작권 보호를 고려하여 해당 데이터는 깃허브나 외부 공개 저장소에는 업로드하지 않았으며, 모든 데이터 처리는 비공개 로컬 환경에서만 수행하였다.

3. 데이터 전처리 과정

3.1 장면 분할 및 영상 데이터 구성

원본 영화 영상은 PySceneDetect라이브러리를 활용하여 장면(Scene) 단위로 자동 분할하였으며, 총 988개의 짧은 영상 클립으로 구성된 장면 데이터셋을 생성하였다. 또한 이미지 기반 장면 이해를 위해, 각 장면 영상에서 대표 프레임을 추출하여 썸네일 이미지를 생성하였고, 해당 이미지는 이후 이미지 캡셔닝 및 시각적 의미 표현 단계에 사용되었다.

3.2 음성 데이터

장면별 음성 정보는 Whisper large-v3 모델을 활용하여 음성-텍스트 변환(STT)을 수행하였다. 이를

통해 각 장면에 대한 대사 및 음성 기반 텍스트 정보를 확보하였으며, 무음 구간이나 의미 없는 음성은 후처리 과정에서 일부 제거하였다. 추가적으로, 각 장면의 음성 파일에 대해 Audio Spectrogram Transformer(AST) 기반 오디오 분류 모델을 적용하여 배경음을 추출하고, 음악, 대화(Speech), 환경음 등 주요 음향 요소를 라벨링하였다.

3.3 등장 인물 데이터 수집 및 라벨링

장면 검색에서 인물 정보를 직접적으로 활용하기 위해, 영화 기생충의 주요 등장인물에 대한 얼굴 데이터와 메타 정보를 별도로 구축하였다. 주연 배우들의 얼굴 이미지, 인물 이름, 배역 이름 정보는 영화진흥위원회(KOFIC)에서 크롤링하여 수집하였으며, 이를 인물별 데이터셋으로 정리하였다. 수집된 얼굴 이미지는 InsightFace를 이용해 얼굴 임베딩 벡터로 변환하였고, 각 벡터는 인물 라벨과 함께 저장되었다. 이후 장면 썸네일 이미지에 대해 얼굴 인식을 수행하여 해당 장면에 등장하는 인물을 자동으로 라벨링할 수 있도록 구성하였다.

4. 모델 선정 및 구현 요약

본 프로젝트는 영화 장면을 단순한 시각 정보가 아닌, 의미 단위로 표현하고 검색하기 위해 멀티모달 기반의 모델 구성과 검색 파이프라인을 설계하였다. 이미지, 음성, 인물 정보를 각각 독립적으로 처리한 뒤, 이를 임베딩 공간에서 통합하여 장면 검색을 수행하는 구조를 채택하였다.

4.1 이미지 캡셔닝 모델 선정 및 활용

장면의 시각적 의미를 표현하기 위해 이미지 캡셔닝 모델을 핵심 구성 요소로 사용하였다. 베이스라인에서는 인물 정보를 포함하지 않은 이미지 캡션을 생성하여 장면의 전반적인 시각적 상황만을 표현하도록 구성하였다. 이후 실험군에서는 다음과 같이 캡셔닝 전략을 점진적으로 확장하였다.

실험군 1: 인물 라벨링 정보를 이미지 캡셔닝에 반영하여 등장 인물이 포함된 장면 의미 표현 생성

실험군 2: 인물 정보가 포함된 이미지 캡션과 Whisper 기반 STT 대사, 그리고 배경음 라벨을 결합하여 장면 의미 표현을 확장

실험군 3: 영상 캡셔닝 결과와 STT 대사를 결합하여 시간 흐름을 반영한 장면 의미 표현 생성

이를 통해 장면의 시각적 정보가 점차 관계 중심·상황 중심 의미 표현으로 확장되도록 설계하였다

4.2 음성 인식 모델(STT) 및 대사 정보 활용

장면의 의미를 보다 풍부하게 표현하기 위해 영상에서 추출한 음성 데이터에 대해 Whisper large-v3 모델을 사용하여 음성-텍스트 변환(STT)을 수행하였다. STT를 통해 생성된 대사는 인물 간의 관계, 상황의 긴장도, 사건 발생 여부 등 이미지 정보만으로는 포착하기 어려운 서사적 단서를 제

공하며, 이미지 캡션과 결합되어 장면의 의미 표현을 강화하는 역할을 한다.

4.3 인물 인식 모델 및 얼굴 임베딩

장면 검색에서 "누가 등장하는지"를 직접적으로 활용하기 위해 InsightFace기반의 얼굴 인식 모델을 사용하였다. 사전에 수집한 주요 등장인물의 얼굴 이미지를 임베딩 벡터로 변환하고, 장면 썸네일 이미지에 대해 얼굴 인식을 수행하여 등장 인물 정보를 자동으로 라벨링하였다. 이를 통해 인물 이름과 배역 정보가 이미지 캡셔닝 및 장면 임베딩 과정에 포함되도록 구성하였다.

4.4 임베딩 생성 및 벡터 검색 구조

각 장면에 대해 생성된 멀티모달 정보인 영상 캡션, 이미지 캡션, Whisper 기반 STT 대사, 인물 라벨 텍스트는 다국어 문장 임베딩 모델인 multilingual-e5-large를 사용하여 의미 벡터로 변환된다. 이를 통해 서로 다른 모달리티에서 생성된 정보가 동일한 의미 공간 상에서 비교 가능하도록 구성하였다.

생성된 장면 임베딩 벡터들은 FAISS(Facebook AI Similarity Search)기반 벡터 인덱스로 구성된 Vector Database에 저장된다. 사용자가 입력한 자연어 질의 역시 동일한 multilingual-e5-large모델을 통해 질의 벡터로 변환되며, 질의 벡터와 장면 벡터 간의 코사인 유사도를 기준으로 벡터 기반 유사도 검색을 수행한다.

검색 결과는 유사도가 높은 순으로 정렬되며, 본 프로젝트에서는 검색 결과의 해석 가능성과 비교 평가를 위해 Top-k 장면 후보를 반환하도록 설정하였다. 이를 통해 사용자 질의와 의미적으로 가장 유사한 장면들을 효율적으로 탐색할 수 있도록 하였다

4.5 시스템 구현 환경

- 임베딩 및 모델 추론은 Python 기반으로 구현

- 대규모 연산을 위해 A100 GPU 환경에서 실험 수행

- Hugging Face 라이브러리를 활용한 모델 로딩 및 관리

- 실험 코드 및 결과 정리는 GitHub를 통해 버전 관리

5. 결과 해석 및 인사이트 도출

본 프로젝트는 단순 캡셔닝 정확도 측정이 아니라, 의미 기반 장면 검색 시스템의 유효성을 검증하는 것이 목적이므로 정성 평가 중심의 평가 전략을 채택하였다.

6. 한계점 및 추후 보완사항

본 프로젝트는 데이터의 대중성과 전반적인 시각적 특성을 고려하여 단일 영화(기생충)를 대상으로 실험을 수행하였기 때문에 다양한 장르나 영상 유형에 대한 일반화 가능성에는 한계가 존재한

다. 또한 영화 장면은 명확한 정답 레이블이 존재하지 않기 때문에 Precision, Recall과 같은 정량 지표를 활용한 평가가 제한적이라는 한계가 있다.

추후 연구에서는 다음과 같은 방향으로 보완이 가능하다.

1. 다수의 영화 및 드라마 데이터로 확장하여 장르 및 서사 구조에 따른 성능 차이 분석

2. 사용자 평가 또는 다중 평가자 기반의 보다 체계적인 정성 평가 기준 도입

3. 의미 기반 장면 검색 성능을 비교·분석할 수 있는 벤치마크 데이터셋 및 평가 프레임워크 구축

- 질의–장면 쌍(annotation)을 기반으로 한 기준 세트 정의

- 의미 일치도, 장면 적합도 등을 포함한 정성·준정량 평가 지표 설계

3. Video LLM을 활용한 시간 흐름 기반 장면 이해 및 요약 기능 확장

4. 장면을 독립적인 단위가 아닌, 영화 전체 서사 흐름 속의 연속된 맥락으로 이해하여 전후 장면의 의미 정보를 함께 고려하는 검색 방식으로 확장

이러한 보완을 통해 본 시스템은 영화 아카이브 검색, OTT 내부 장면 탐색, 콘텐츠 분석 등 다양한 응용 분야로 확장 가능할 것으로 기대된다.