JDCS
디지털콘텐츠학회논문지
Journal of Digital Contents Society
Vol. 26, No. 2, pp. 471-478, Feb. 2025

Check for updates

비전 언어 모델을 이용한 영화 태그 생성 기법

김 현 민[1] · 유 준[2*]
[1]가천대학교 AI·소프트웨어학부 학사과정
[2]가천대학교 AI·소프트웨어학부 교수

Movie Tag Generation Method Using Vision-Language Models

Hyun-min Kim[1] · Joon Yoo[2*]

[1]Bachelor's Course, School of Computing, Gachon University, Seongnam 13120, Korea
[2]Professor, School of Computing, Gachon University, Seongnam 13120, Korea

[요 약]

본 연구는 사전학습된 비전-언어 모델을 활용하여 영화 이미지에서 자동으로 태그를 생성하는 기법을 제안한다. 본 연구에서는 LLaVA 모델을 사용하여 영화 프레임별로 정교한 캡션을 생성한 후, 이를 바탕으로 보다 세밀하고 다양한 태그를 할당하는 과정을 수행한다. 특히, 영화 이미지 데이터를 활용해 모델을 파인튜닝 함으로써 태그 생성 성능을 크게 향상시켰다. 이러한 접근법은 콘텐츠의 시각적 특징을 효과적으로 추출하고, 수작업 태깅 과정에서 발생할 수 있는 주관성과 시간 소모 문제를 해결하는 데 기여할 수 있다. 그러나 상업적 활용을 위해서는 저작권 문제와 데이터 신뢰성 문제 등의 한계가 존재하며, 이에 대한 추가적인 연구와 개선이 필요하다. 본 연구는 멀티모달 추천 시스템, 영상 검색, 전자상거래 등 다양한 분야에서 이 방법론이 폭넓게 활용될 가능성을 시사하며, 향후 다양한 응용 분야에서 중요한 기여를 할 수 있을 것으로 기대된다. 또한, 본 연구에서 제안한 태그 생성 방식은 사용자 맞춤형 콘텐츠 제공과 같은 새로운 서비스 개발에도 기여할 수 있을 것이다.

[Abstract]

This study proposes a method for automatically generating tags from movie images using pre-trained vision-language models. The LLaVA model was employed to generate detailed captions for individual movie frames, which were subsequently utilized to assign more nuanced and diverse tags. Notably, the model was fine-tuned with movie image data, resulting in a significant improvement in tag generation performance. This approach demonstrates its capability to effectively extract visual features from content, addressing the subjectivity and time-intensive nature of manual tagging. Nevertheless, limitations related to copyright issues and data reliability must be resolved to enable commercial applications, necessitating further research and refinement. This study highlights the potential for broad applications of this methodology across various fields, including multimodal recommendation systems, video search, and e-commerce, underscoring its capacity for significant contributions in future applications. Furthermore, the proposed tag generation method could support the development of new services, such as personalized content delivery, thereby enhancing the user experience.

색인어 : 영화, 태그, 대형언어모델(LLM), 비전모델, 이미지
Keyword : Movie, Tag, Large Language Model (LLM), Vison Model, Image

http://dx.doi.org/10.9728/dcs.2025.26.2.471

Received 22 November 2024; **Revised** 24 December 2024
Accepted 14 January 2025

***Corresponding Author; Joon Yoo**

Tel: +82-31-750-5832
E-mail: joon.yoo@gachon.ac.kr

Ⅰ. 서 론

멀티모달 대규모 언어모델, 그 중에서도 비전 언어 모델은 여러가지 SOTA (State of the art) 모델을 거치며 비전-언어 멀티모달(Vision-language Multimodal) 분야에서 눈부신 활약을 보여주었다. 예를 들어, ViLBERT[1]는 시각적-언어적 표현을 사전 학습하여 다양한 비전-언어 태스크에서 뛰어난 성능을 입증하였으며 VisualBERT[2]는 간단하면서도 강력한 성능을 보이는 모델로, 비전-언어 연계 태스크의 기초 기준을 제시하였다. 또한 CLIP[3] 모델은 자연어 감독(Supervision) 하에 학습된 시각적 모델로, 다수의 비전-언어 작업에서 탁월한 성능을 보여주며 큰 주목을 받았다. 이러한 모델들은 비전-언어 멀티모달 분야의 발전을 이끌며, 다양한 응용 분야에서 활용되고 있다. 비전-언어 모델을 통해 이미지 캡셔닝과 질의응답을 비롯한 여러가지 비전 태스크를 수행할 수 있게 되었고, 텍스트 기반 이미지 생성, 이미지 내 피처(feature) 분석 등 다양한 애플리케이션에서 이용되고 있다. 특히 VQA (Visual Question Answering)[4] 모델은 주어진 이미지에 대한 질문에 답변할 수 있는 모델로, 비전-언어 모델이 다양한 비전 태스크에서 강력한 성능을 보이는 대표적인 사례 중 하나이다.

사람들의 여가 시간이 늘어나고, 영화를 시청하는 것을 좋아하는 사람들이 늘어남에 따라 다양한 OTT (Over The Top) 서비스가 확대되고 있는 추세이다. 특히 그 중 대표격인 '넷플릭스' 플랫폼에서는 사용자들이 흥미를 느낄 만한 컨텐츠를 상단에 노출하기 위해 추천 알고리즘 대회인 “NETFLIX PRIZE”도 개최할 만큼 다양한 노력을 기울이고 있다[5]. 이러한 노력은 넷플릭스가 사용자 맞춤형 경험을 제공하는 데 큰 역할을 했다. 이와 같이 컨텐츠를 추천하기 위해서나, 정보를 한눈에 알기 위해서는 컨텐츠의 특징을 데이터로 나타낸 정보가 필요하다. 영화를 예로 들면 주연 배우나 장르, 본 논문에서 다룰 태그 정보들이 이에 해당한다.

하지만 컨텐츠에서 이러한 피처들을 추출하는 데는 많은 시간과 비용이 발생하고, 태그를 작성하는 태거에 따라 주관적인 기준으로 작성될 수 있다는 문제가 있다.

본 연구에서는 비전-언어 모델을 이용한 이미지 캡셔닝 태스크(Captioning task)를 기반으로 효과적인 영화 태그 생성을 목표로한다. 기존의 영화 태깅 방식과는 달리, 본 연구에서는 Multimodal LLM (Large Language Model)의 정교한 이미지 캡셔닝을 이용하여 영화 태그를 생성하는 것이 가능한지 분석하고, 이를 활용할 수 있는지에 대한 가능성을 제안한다. 이를 위해 영화 제목 키워드 기반 유튜브 크롤링을 통해 수집된 영화 예고편, 리뷰 컨텐츠 및 명장명 숏폼 영상 등 영화 관련 영상 및 프레임별로 분할된 이미지를 통해 비전-언어 모델을 활용하여 영화 태그를 생성하고, 영화 속 이미지를 통해 파인튜닝 된 모델을 통해 태그 생성의 성능을 높이는 방법을 보여주었으며, 향후 다양한 응용 가능성을 제시한다.

이후 본 논문은 다음과 같이 구성되어 있다. 2장에서는 비전-언어 모델 및 기존 태깅 방법에 관한 관련 연구를 다루며, 3장에서는 본 연구에서 제시하는 방법론과 전반적인 연구 환경 및 연구 진행 과정을 소개한다. 4장에서는 연구 결과를 분석하며 비교하고 본 연구에서 나타난 한계를 분석한다. 5장에서는 결론과 함께 본 연구의 다양한 활용 가능성을 제시한다.

Ⅱ. 관련 연구

2-1 Vision-Language Pretrained model

비전 언어 모델(VLP; Vision-Language Pretrained Models)은 시각적 정보와 텍스트 정보를 결합하여 대규모 데이터셋을 이용하여 사전 학습된 모델로, 시각적인 특징과 언어적인 특징과 관련된 여러가지 표현을 학습한다. 예를 들어 CLIP 모델은 다양한 Image-Text Pair를 학습하여 이미지를 텍스트 설명과 연결하는 방법을 학습하고, ALIGN[6]은 CLIP에서 모델 크기와 학습 데이터를 확장하여 성능을 향상시켰다. 그리고 본 연구에서 사용한 LLaVA 모델의 경우 Vicuna 모델로 초기화된 파라미터를 CLIP 비전 인코더와 연결하여 더욱 이미지를 잘 이해하고 원활한 대화가 가능하도록 학습된 모델이다[7]. 이와 같은 VLP 모델은 이미지 캡셔닝, 이미지 분류, 시각적 질의 응답(VQA; Visual Question Answering)과 같은 작업에서 좋은 성능을 보여준다.

2-2 영화 태그 생성

영화의 태그는 영화의 정보를 표시하기 위해 붙이는 정보로, 영화의 분위기나 장르, 간단한 설명을 제공한다. 예를 들어, 영화 Inception의 경우 ‘Sci-Fi’, ‘Thriller’, ‘Mind-Bending’과 같은 태그가 붙을 수 있다. 이러한 태그는 영화의 장르와 분위기를 반영하지만, 종종 Tagger들의 주관적 판단에 따라 결정된다. 기존 방법들은 태그를 작성하는 Tagger들이 수동으로 태깅하거나, 맥락과 뉘앙스가 부족한 간단한 시놉시스 키워드 추출 및 사용자 리뷰를 이용한 방식에 의존했다[8]. 최근에는 이러한 한계를 극복하기 위해 기계학습을 이용한 자동 태그 생성 방법이 제안되었으며, 이는 영화 장르 분류 및 추천 시스템에 긍정적인 영향을 미치고 있다[9]. 최근 연구에서는 영화의 오디오 피처를 활용하여 자동 태그를 생성하는 방법이 제안되었다. 이 방법은 배경음과 오디오 신호를 분석하고 이를 태그 예측 모델에 적용함으로써 태그 생성의 정확도를 높이는 방식으로, 영화 추천 시스템의 성능을 개선하는 데 기여하고 있다.

MPST (Movie Plot Synopses with Tags)[10] 논문에서는 위키피디아(Wikipedia)나 IMDb (Internet Movie Database)에서 수집한 영화의 줄거리 시놉시스 데이터를 기반으로 태그를 생성하는 방법을 제시하였다. MPST 데이터

셋은 14,000개 이상의 영화 제목 및 영화에 대한 줄거리 요약 데이터와 71개의 고유 태그를 포함하며, 위키피디아[11]는 사용자들이 편집한 정보를 포함하는 오픈 백과사전으로 영화의 줄거리나 제작 정보 등을 제공한다. IMDb[12]는 영화의 줄거리, 출연진, 평점 등을 포함한 세계 최대의 영화 데이터베이스이다.

Ⅲ. 방법론

3-1 데이터 수집

영화의 이미지 정보를 이용하여 태깅하는데 사용하려면 영화의 영상 데이터가 필요하다. 하지만 대부분의 영화의 경우 저작권 문제로 무료로 영상 데이터를 얻을 수 없기에 본 연구에서는 웹 크롤링을 통해 유튜브 영상 데이터를 수집한다. 영화 제목 키워드를 기반으로 유튜브 크롤링을 실행하면 영상의 제목과 URL (Uniform Resource Locator) 등의 메타데이터를 얻을 수 있다. 영화 제목으로 검색 시 상단에 노출되는 영상들의 경우 영화의 예고편이나 다양한 영화 유튜버들의 리뷰 영상, 영화 내 명장면 숏 폼 영상 등이 포함되며, 이런 영상들의 경우 영화의 다양한 장면을 설명하고 태그를 생성하기에 적절하다고 판단하여 데이터 소스로 사용하게 되었다. URL을 통해 영상을 다운로드하고 프레임 단위로 추출하여 이미지 데이터를 얻은 뒤 Vision-Language 모델을 통해 이미지 속 장면을 기반으로 태그를 생성하게 된다.

3-2 태그 데이터 테이블 생성

비전-언어 모델을 통해 얻은 프롬프트에서 태그만 추출하기 위해서는 태그 데이터 테이블이 필요하다. 모델에게 태그에 대한 정보를 요청하여 얻을 수 있는 프롬프트는 List, json 형태로 반환하거나 긴 문장으로 설명해주는 등 요청 시마다 다른 형태로 얻을 수 있기에 모델이 생성한 설명에 미리 준비한 태그 테이블의 어휘가 포함되어 있는지 확인하는 방식으로 태그 정보를 생성하는 방법을 사용했다. 먼저 MPST 데이터셋에 포함된 고유 태그를 테이블로 만들어 표 1과 같이 저장하였고, 71개의 고유 태그를 얻을 수 있었다. 이후 모델의 프롬프트에 고유 태그 feature의 단어가 포함되는지 대응하여 해당 키워드의 영화에 태그를 할당하였다. 해당 데이터셋의 태그 중 고유 태그 테이블에 포함된 태그를 활용하여 파인튜닝에 필요한 프롬프트를 작성하였다.

3-3 모델을 이용하여 태그 생성

프레임별로 생성된 영화 이미지의 경우 영화 속 다양한 장면을 각각 포함하고 있어, 각 이미지로는 영화의 전반적인 태그를 생성하기 어려울 것이다. 액션 영화나 범죄 영화에 로맨틱한 장면이 포함되거나, 반대로 멜로 영화에서 폭력적인 장면이 등장하는 경우 등 여러가지 예시를 떠올릴 수 있다.

표 1. 고유 태그 테이블
Table 1. Unique tags table

Index	Unique Tags
1	cult
2	horror
3	gothic
4	murder
…	…
71	non fiction

따라서 각각의 이미지에서 추출한 태그를 개수별로 해당 영화 테이블에 나열하여 태그 리스트를 생성한다. 이후 Counting을 통해 해당 영화에 태그를 할당하는 데 사용된 전체 이미지 수와 대비하여 태그가 등장한 비율을 나타내도록 한다. 전체 진행 과정의 알고리즘은 그림 1과 같다.

Algorithm 1 MainProcess

```
Input: searchQueries, maxResults,baseDir = "videos"
Output: Downloaded video, extracted frames, CSV results
// 1) YouTube crawling
for each query in searchQueries do
    movieDir ← join(baseDir, query)
    results ← CrawlYouTubeVideos(query, maxResults)
    for each res in results do
        DownloadYouTubeVideo(res.url, movieDir)
    end for
end for
// 2) Extract frames per video
for each query in searchQueries do
    movieDir ← join(baseDir, query)
    if ¬exists(movieDir) then
        continue
    end if
    videoFiles ← all .mp4 under movieDir
    for each vidFile in videoFiles do
        vidPath ← join(movieDir, vidFile)
        SaveFrames(vidPath, 640, 360, movieDir)
    end for
end for
// 3) Analyze images and save to CSV
AnalyzeImagesAndSaveResults(baseDir,
"movie_tags_results.csv")
end algorithm
```

그림 1. 태그 생성 과정 알고리즘
Fig. 1. Tag generation process algorithm

3-4 비전 언어 모델 파인튜닝

프레임별로 나누어진 영화 이미지와 태그 정보를 이용한 프롬프트를 통해 모델을 파인튜닝하여 보다 섬세하고 정확한 태그를 생성하도록 트레이닝한다. 프롬프트에 사용된 태그 정보는 MovieLens Tag-genome 데이터셋을 활용하였다. MovieLens Tag-genome 데이터셋은 유저 어플리케이션, 텍스트 리뷰, 평가 등 User-generated content를 기반으로

제작한 데이터셋이며, 사용자의 설문을 바탕으로 학습된 모델이 생성한 태그와 영화의 relevance (genome-score) 점수를 포함한 데이터셋이다[13]. Google colab에서 제공하는 A100 GPU를 통해 파인튜닝을 시도했다. 파인튜닝에 사용된 영화 이미지는 MPST 데이터셋의 상위 101번째부터 200번째까지 영화 타이틀 중 20개의 샘플 영화를 사용하여 진행했고, 총 864장의 분할된 이미지를 이용하였다. 프롬프트는 MovieLens Tag-genome 데이터셋에서 Unique tags 테이블에 있는 태그를 이용하여 각 영화마다 통일하였다.

(Ex. “17 Again” 영화 이미지의 프롬프트 : “A scene from a movie containing tag data ‘comedy’, ‘feel-good’, ‘entertaining’, ‘romantic’, ‘cute’, ‘storytelling’, ‘clever’, ‘absurd’, ‘dramatic’, ‘stupid’, ‘fantasy’, ‘alternate reality’, ‘sentimental’, ‘revenge’, ‘boring’, ‘brainwashing’, ‘action’, ‘plot twist’, ‘queer’, ‘comic’, ‘whimsical’, ‘thought-provoking’, ‘good versus evil’, ‘atmospheric’, ‘alternate history’, ‘humor’, ‘murder’, ‘allegory’, ‘psychological’, ‘suspenseful’, ‘philosophical’, ‘violence’, ‘gothic’, ‘historical’, ‘insanity’, ‘tragedy’, ‘satire’, ‘sci-fi’, ‘inspiring’, ‘mystery’, ‘bleak’, ‘claustrophobic’, ‘cult’, ‘dark’, ‘horror’, ‘western’, ‘blaxploitation’, ‘depressing’, ‘psychedelic’”)

5회의 epoch를 거쳐 트레이닝 하였으며 LLaVA 전체 모델과 파인튜닝 된 LoRA weight를 합하여 최종 모델을 얻을 수 있었다. A100 GPU를 통해 트레이닝 하였으며 트레이닝에는 약 12분가량 소요되었다. 이후 기존 사전학습 된 모델(VLP)을 사용한 태깅과 파인튜닝 된 모델의 태그 할당 결과를 비교하여 유의미한 결과 차이가 있는지 비교한다.

Ⅳ. 실험 및 실험 결과

4-1 연구 환경과 결과 분석

MPST 데이터셋과 태깅 결과를 비교하기 위해 MPST 데이터셋의 상위 100개 영화 타이틀을 이용하여 태깅을 진행했다. 사용한 하드웨어는 Titan XP GPU 2개로 사전학습 된 LLaVA 모델을 이용하여 이미지 캡셔닝 및 프롬프트를 생성하였다. LLaVA 모델은 기존 대규모 언어 모델이 가진 문맥 이해·생성 능력에 시각적 이해가 추가되어, 보다 폭 넓은 문제를 해결할 수 있는 모델이다[7]. 유튜브 키워드 검색 결과 상위 5개의 영상이 모두 이용할 수 없는 영상인 경우를 제외하고 63개 영화에 대한 태그를 생성하였다. 영상을 더 폭넓게 가져와 분석한다면 상위 검색 결과가 이용할 수 없는 영상인 경우에도 태그를 생성할 수 있지만, 태그 생성 시간이 길어지고 연관 없는 영상이 데이터셋에 포함될 가능성이 늘어날 것으로 추정한다. 연구에는 총 3139개의 이미지, 영화 당 평균 49.8개의 이미지가 사용되었다. 모델이 전체 이미지를 분석하고 프롬프트를 작성하는 데는 약 28.7시간이 소요되었다.

4-2 VLP 모델 태그 할당 결과

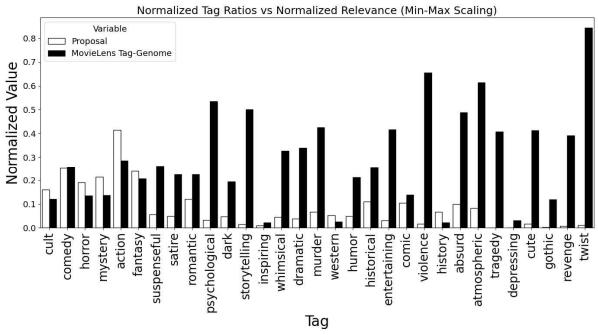

그림 2. 사전학습 비전 언어 모델 태그 할당 결과
Fig. 2. VLP model tag assignment results

MPST (MPST : Movie Plot Synopses with Tags) 데이터셋은 IMDb와 위키피디아의 영화 줄거리 시놉시스를 이용하여 태그를 추출한 데이터셋이다. MPST 데이터셋의 경우 14828개의 영화에 태그를 할당하였으며 영화 하나 당 평균 태그 개수는 약 2.98개이다. 본 연구에서 VLP모델을 통해 MPST의 상위 100개 영화에서 얻은 평균 태그 수는 약 30개로, MPST보다 매우 많은 태그를 통해 영화를 더 풍부하게 설명할 수 있다. 또한 생성된 태그의 품질을 알아보기 위해 MovieLens Tag-genome 데이터셋과 비교했다. 시각화 그래프를 보면 “cult, comedy, action, fantasy”와 같은 태그들은 relevance 점수와 비슷한 연관성 수치로 태깅이 잘 되었다고 볼 수 있다. 하지만 “psychological”과 같이 relevance에 비해 많이 생성되지 않은 태그들도 있다. 그 이유로는 첫번째로 일반적으로 사용되는 단어 위주로 모델이 프롬프트를 생성했기 때문으로 추측된다. 모델 쿼리가 길어질수록 모델의 추론 시간이 길어지므로 태그 테이블에 있는 단어를 모두 쿼리에 넣지 않았기 때문에, 자주 사용되지 않는 단어의 경우 관련된 태그를 생성하지 않았기 때문이다. 두번째로 “storytelling”과 같이 영화의 진행 방식을 나타내는 태그나, “non fiction”과 같이 분할된 이미지를 통해서는 알아내기 힘든 태그 정보의 경우 비전 언어 모델만을 이용해서는 이와 같은 정보를 얻기 힘들기 때문으로 추정한다.

4-3 모델 파인튜닝 결과

파인튜닝 전 모델 태깅 결과와 비교해 전체적인 태그 생성률이 늘어났다. 그에 따라 태그 단어들이 문장에 등장하는 빈도가 늘어났고, 이전 사전학습 모델을 사용했을 때에 비해

표 2. 파인튜닝 된 비전 언어 모델을 통한 태깅 결과 예시
Table 2. Example of tagging results using a fine-tuned vision language model

Movie	Tags (count, Appearance ratio)	Image count
Sinbad and the Eye of the Tiger	gothic(63, 76.83%), plot twist(68, 82.93%), psychedelic(43, 52.44%), dark(68, 82.93%), comic(58, 70.73%), humor(56, 68.29%), insanity(57, 69.51%), sci-fi(46, 56.10%), whimsical(57, 69.51%), queer(52, 63.41%), dramatic(56, 68.29%), fantasy(56, 68.29%), brainwashing(57, 69.51%), inspiring(26, 31.71%), claustrophobic(61, 74.39%), historical(72, 87.80%), stupid(53, 64.63%), philosophical(23, 28.05%), comedy(52, 63.41%), psychological(61, 74.39%), alternate reality(43, 52.44%), alternate history(45, 54.88%), feel-good(40, 48.78%), cute(53, 64.63%), historical fiction(7, 8.54%), tragedy(32, 39.02%), allegory(52, 63.41%), revenge(66, 80.49%), horror(59, 71.95%), action(74, 90.24%), atmospheric(59, 71.95%), murder(43, 52.44%), cult(52, 63.41%), bleak(55, 67.07%), depressing(45, 54.88%), suspenseful(68, 82.93%), mystery(34, 41.46%), violence(58, 70.73%), good versus evil(71, 86.59%), storytelling(70, 85.37%), romantic(67, 81.71%), boring(23, 28.05%), entertaining(45, 54.88%), absurd(27, 32.93%), thought-provoking(33, 40.24%), satire(31, 37.80%), sentimental(34, 41.46%), clever(35, 42.68%), western(37, 45.12%), blaxploitation(28, 34.15%)	82
Little Caesar	dark(48, 56.47%), dramatic(29, 34.12%), fantasy(27, 31.76%), claustrophobic(36, 42.35%), comedy(34, 40.00%), horror(30, 35.29%), western(30, 35.29%), cult(27, 31.76%), suspenseful(39, 45.88%), gothic(33, 38.82%), plot twist(37, 43.53%), psychedelic(28, 32.94%), boring(17, 20.00%), entertaining(20, 23.53%), comic(34, 40.00%), humor(30, 35.29%), insanity(31, 36.47%), whimsical(32, 37.65%), queer(31, 36.47%), absurd(17, 20.00%), historical(38, 44.71%), stupid(30, 35.29%), philosophical(13, 15.29%), psychological(37, 43.53%), alternate reality(19, 22.35%), alternate history(29, 34.12%), feel-good(27, 31.76%), cute(31, 36.47%), thought-provoking(23, 27.06%), tragedy(21, 24.71%), allegory(25, 29.41%), revenge(31, 36.47%), action(34, 40.00%), atmospheric(35, 41.18%), murder(24, 28.24%), bleak(34, 40.00%), depressing(32, 37.65%), mystery(24, 28.24%), sentimental(36, 42.35%), violence(26, 30.59%), good versus evil(34, 40.00%), storytelling(39, 45.88%), romantic(37, 43.53%), sci-fi(25, 29.41%), blaxploitation(23, 27.06%), brainwashing(29, 34.12%), inspiring(17, 20.00%), clever(19, 22.35%), satire(14, 16.47%), historical fiction(9, 10.59%)	85
Seven Years in Tibet	dark(65, 67.71%), gothic(49, 51.04%), plot twist(51, 53.12%), comic(42, 43.75%), humor(46, 47.92%), insanity(43, 44.79%), whimsical(43, 44.79%), dramatic(58, 60.42%), brainwashing(44, 45.83%), claustrophobic(52, 54.17%), historical(65, 67.71%), stupid(39, 40.62%), comedy(40, 41.67%), psychological(49, 51.04%), alternate history(36, 37.50%), cute(45, 46.88%), revenge(55, 57.29%), horror(44, 45.83%), action(53, 55.21%), western(47, 48.96%), cult(45, 46.88%), bleak(51, 53.12%), depressing(46, 47.92%), suspenseful(53, 55.21%), sentimental(38, 39.58%), good versus evil(64, 66.67%), storytelling(66, 68.75%), romantic(59, 61.46%), fantasy(27, 28.12%), murder(29, 30.21%), queer(39, 40.62%), absurd(25, 26.04%), feel-good(34, 35.42%), historical fiction(11, 11.46%), allegory(42, 43.75%), atmospheric(51, 53.12%), violence(44, 45.83%), psychedelic(24, 25.00%), boring(12, 12.50%), entertaining(26, 27.08%), philosophical(18, 18.75%), alternate reality(24, 25.00%), thought-provoking(27, 28.12%), tragedy(24, 25.00%), mystery(26, 27.08%), clever(30, 31.25%), sci-fi(25, 26.04%), inspiring(18, 18.75%), blaxploitation(22, 22.92%), satire(18, 18.75%)	96
Flightplan	gothic(49, 53.85%), plot twist(53, 58.24%), psychedelic(33, 36.26%), boring(16, 17.58%), entertaining(30, 32.97%), dark(65, 71.43%), comic(48, 52.75%), humor(46, 50.55%), insanity(45, 49.45%), sci-fi(37, 40.66%), whimsical(42, 46.15%), queer(43, 47.25%), dramatic(46, 50.55%), fantasy(34, 37.36%), brainwashing(48, 52.75%), absurd(29, 31.87%), inspiring(17, 18.68%), claustrophobic(51, 56.04%), blaxploitation(30, 32.97%), historical(49, 53.85%), stupid(40, 43.96%), philosophical(20, 21.98%), comedy(43, 47.25%), psychological(56, 61.54%), alternate reality(28, 30.77%), alternate history(32, 35.16%), feel-good(32, 35.16%), cute(43, 47.25%), thought-provoking(27, 29.67%), tragedy(24, 26.37%), allegory(34, 37.36%), revenge(50, 54.95%), horror(44, 48.35%), action(54, 59.34%), atmospheric(46, 50.55%), murder(27, 29.67%), western(34, 37.36%), cult(35, 38.46%), bleak(43, 47.25%), depressing(46, 50.55%), satire(26, 28.57%), suspenseful(57, 62.64%), mystery(40, 43.96%), sentimental(43, 47.25%), clever(31, 34.07%), violence(40, 43.96%), good versus evil(50, 54.95%), storytelling(58, 63.74%), romantic(50, 54.95%), historical fiction(2, 2.20%)	91
…	…	…
Big Nothing	gothic(33, 38.37%), plot twist(33, 38.37%), psychedelic(25, 29.07%), boring(12, 13.95%), entertaining(16, 18.60%), dark(43, 50.00%), comic(35, 40.70%), humor(34, 39.53%), insanity(31, 36.05%), sci-fi(26, 30.23%), whimsical(28, 32.56%), queer(22, 25.58%), dramatic(28, 32.56%), fantasy(28, 32.56%), brainwashing(31, 36.05%), absurd(16, 18.60%), inspiring(9, 10.47%), claustrophobic(34, 39.53%), blaxploitation(20, 23.26%), historical(35, 40.70%), stupid(30, 34.88%), philosophical(13, 15.12%), comedy(32, 37.21%), psychological(35, 40.70%), alternate reality(16, 18.60%), alternate history(22, 25.58%), feel-good(23, 26.74%), cute(36, 41.86%), thought-provoking(17, 19.77%), tragedy(13, 15.12%), allegory(23, 26.74%), revenge(28, 32.56%), horror(30, 34.88%), action(36, 41.86%), atmospheric(29, 33.72%), murder(15, 17.44%), western(24, 27.91%), cult(22, 25.58%), bleak(33, 38.37%), depressing(31, 36.05%), satire(21, 24.42%), suspenseful(35, 40.70%), mystery(23, 26.74%), sentimental(30, 34.88%), clever(16, 18.60%), violence(23, 26.74%), good versus evil(34, 39.53%), storytelling(34, 39.53%), romantic(35, 40.70%), historical fiction(3, 3.49%)	86

Unique tags 테이블에 있는 잘 사용되지 않는 단어들도 비교적 잘 생성된 것을 볼 수 있다. Accuracy 측정 방법은 기존 태그 대비 알맞게 할당된 태그의 비율인 태그 매칭률 (match ratio)로 환산하였다.

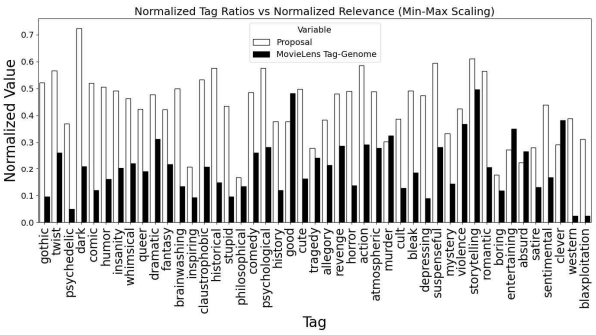

그림 3. 파인튜닝 후 비전 언어 모델 태그 할당 결과
Fig. 3. Model tag assignment results after finetuning

MovieLens tag-genome 데이터셋의 태그 중 고유 태그 테이블에 있는 태그 대비 알맞게 할당된 태그 비율을 그림 4, 5와 같이 얻을 수 있었다. 또한 MovieLens tag-genome 데이터셋과 비교하여 잘못 할당된 태그의 평균 개수 및 모델이 찾아내지 못한 태그 개수를 표 3과 같이 확인해볼 수 있다.

표 3. 파인튜닝 전후 태그 할당 결과 비교
Table 3. Comparison of tag assignment results

	VLP	Finetuning
Accuracy	28.88%	**89.84%**
Average Unmatched Tags	9.56	**5.04**
Average Missing Tags	13.33	**4.00**

모델이 생성하는 프롬프트에 태그 관련 단어들이 많이 등장하게 됨에 따라 평균 태그 매칭률은 3배가량 증가하였으며 위 방법으로 찾지 못한 Missing tag의 개수도 확연하게 감소하였다. 그럼에도 여전히 “fiction, reality, plot twist” 등과 같이 분할된 이미지만으로 확인하기 어려운 태그들의 경우 비전 언어 모델로는 알맞게 할당하는 데 한계가 있다는 점을 확인할 수 있었다. 하지만 사용자 정보와 평가, 설문을 이용하는 과정 없이도 자주 쓰이는 태그들을 효과적으로 생성할 수 있다는 점에서 의의가 있다.

4-4 Ablation Study

사용자 프롬프트의 복잡도를 감소시키고 LoRA 랭크 (r)을 조정하여 모델 학습 성능에 미치는 영향을 평가하였다.[14] 사용자 프롬프트의 태그 개수를 전체 태그에서 3개로 축소하였으며, LoRA 랭크를 각각 64, 128, 256으로 설정하였고 이외 하이퍼파라미터는 동일하게 유지하였다. 또한 이미지 크기에 따른 변화를 평가하기 위해 이미지 해상도에 따른 train loss를 각각 원본 사이즈, 360x180, 180x90으로 변경하여 학습하였다. 학습 결과 LoRA 랭크가 증가할수록 train loss가 지속적으로 감소하는 경향을 확인할 수 있었다. 각각 r = 64에서 0.1940, r = 256에서 0.0818로 가장 낮은 train loss를 기록하였다. 그러나 r = 256의 경우 학습 도중 기울기 폭주(Gradient exploding)가 발생하여 학습 안정성에 문제가 있음을 확인하였다. 이미지 해상도 별 train loss는 표 4와 같이 나타났다. 원본 크기를 사용했을 때 train loss는 0.0814로 가장 낮았으며, 축소될수록 소폭 증가하였다. 이는 해상도가 감소함에 따라 일부 정보가 손실될 수는 있으나, 학습 성능에 큰 영향을 미치지 않는 모습을 보여준다.

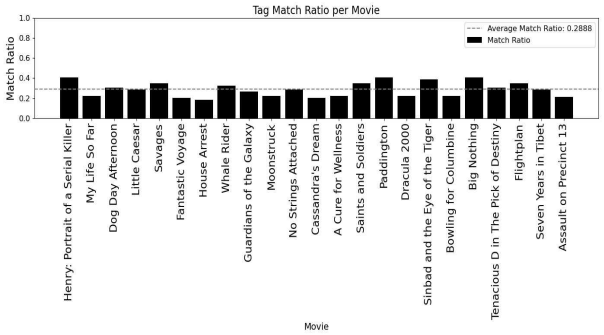

그림 4. 파인튜닝 전 비전 언어 모델 태그 할당률
Fig. 4. VLP model tag match ratio

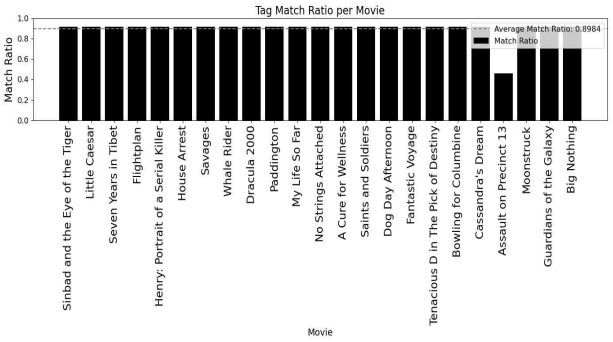

그림 5. 파인튜닝 후 비전 언어 모델 태그 할당률
Fig. 5. Vision language model tag match ratio after finetuning

표 4. 이미지 해상도에 따른 train loss 비교
Table 4. Comparison of train loss according to image size

Image size	Train loss
original size	0.0814
360x180	0.0815
180x90	0.0818

4-5 한계

1) 데이터 수집 관련 저작권 문제

태그 분석에 사용된 영화 이미지 데이터는 주로 유튜브에

업로드 된 다양한 리뷰 유튜버들의 리뷰 영상이나 숏폼 편집 영상, 그리고 영화 예고편이 사용된다. 이러한 영상들의 저작권은 영화사에 있으며, 각 영화의 2차 창작물의 경우 영상을 편집하고 제작한 업로더에게 있다. 본 연구는 이러한 저작물을 통하여 이미지를 추출, VLP 모델을 통해 태그를 생성하는 방법론을 제안하기 위하여 진행되었지만 웹 영상 저작물을 사용하여 진행되었으므로 결과물을 상업용으로 사용하거나 수익을 얻기 위한 어플리케이션 제작을 위해 사용한다면 저작권 문제가 발생할 가능성이 있다.

2) 검증되지 않은 영화 정보

본 연구에서는 데이터를 수집하기 위해서 영화 제목을 키워드 기반으로 유튜브 웹 크롤링을 수행했다. 유튜브에는 수많은 영상들이 있고, 다양한 영화에 대한 정보가 넘쳐나기 때문이다. 하지만 인지도가 아주 낮은 영화의 경우 키워드를 기반으로 검색했을 때 해당 영화에 대한 예고편이나 리뷰 영상 등이 유튜브에 존재하는지 확인하기 어렵고, 또 운이 나쁜 경우 비슷한 제목의 영상이 대신 상단에 노출되는 등 적절한 영상이 태그 생성에 사용되었는지 알 수 없다는 한계가 있다.

V. 결 론

본 연구에서는 비전-언어 모델을 활용하여 영화 이미지에서 자동으로 태그를 생성하는 방법을 제안하였다. 사전학습된 LLaVA 모델을 사용하여 프레임 단위로 추출된 영화 이미지에 대한 캡션을 생성하고, 이를 기반으로 태그를 할당하는 방식을 통해 더 풍부하고 세밀한 태그를 생성할 수 있었다. 또한 분할된 영화 이미지와 태그 정보를 포함한 프롬프트를 이용한 파인튜닝을 통해 태그 모델을 조정하여 성능을 향상했다. 본 연구를 통해 VLP 기반 모델이 태깅 작업에서도 유의미한 파인튜닝 성능을 발휘함을 입증하였다. 특정 영화 장면에 대한 태깅 작업에서 높은 정밀도를 보였으며, 기존 사전학습 모델과 비교하여 우수한 성능을 나타내었다. 생성된 태깅 결과는 MPST 데이터셋의 태그와 비교했을 때도 더 풍부한 태그를 생성할 수 있음을 보였으며, 이는 MPST 데이터셋과 MovieLens Tag-genome과는 다르게 멀티모달 학습 환경에서 사용자 데이터 없이도 제안된 접근법이 효과적임을 시사한다. 향후 태그의 다양성과 사용자 정의 태그의 생성 가능성을 확대하기 위해 동적인 프롬프트 생성을 도입하는 등의 방법으로 본 연구에서 제안된 접근법이 더 폭넓게 활용될 수 있을 것으로 기대된다. 한계점으로는 저작권 문제와 데이터 신뢰성 문제가 있었다. 상업적 용도를 위해서는 저작권 검토가 필요하며, 데이터 수집 과정에서 필터링 기법의 개선이 필요하다. 본 연구는 비전-언어 모델을 활용한 영화 태그 생성의 가능성을 제시함으로써, 향후 멀티모달 추천 시스템에 기여할 수 있을 것이다. 또한 이러한 접근법은 태그 기반 영상 검색이나 쇼핑 플랫폼에서 상품 이미지 기반 태그 검색 등 다양한 분야에도 활용될 수 있을 것이다.

감사의 글

본 연구는 2021년도 과학기술정보통신부 이공분야기초사업의 지원(NRF-2021R1F1A1063640) 및 2020년도 과학기술정보통신부 및 정보통신기획평가원의 SW중심대학사업 지원을 받아 수행하였음.

참고문헌

[1] J. Lu, D. Batra, D. Parikh, and S. Lee, “ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks,” in *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, Vancouver, Canada, pp. 13-23, December 2019. https://doi.org/10.48550/arXiv.1908.02265

[2] L. H. Li, M. Yatskar, D. Yin, C.-J. Hsieh, and K.-W. Chang, “VisualBERT: A Simple and Performant Baseline for Vision and Language,” arXiv:1908.03557, August 2019. https://doi.org/10.48550/arXiv.1908.03557

[3] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, ... and I. Sutskever, “Learning Transferable Visual Models From Natural Language Supervision,” in *Proceedings of the 38th International Conference on Machine Learning*, Online, pp. 8748-8763, July 2021. https://doi.org/10.48550/arXiv.2103.00020

[4] A. Agrawal, J. Lu, S. Antol, M. Mitchell, C. L. Zitnick, D. Parikh, and D. Batra, “VQA: Visual Question Answering,” arXiv:1505.00468v1, May 2015. https://doi.org/10.48550/arXiv.1505.00468

[5] J. Bennett and S. Lanning, “The Netflix Prize,” in *Proceedings of KDD Cup and Workshop 2007*, San Jose: CA, pp. 1-4, August 2007.

[6] C. Jia, Y. Yang, Y. Xia, Y.-T. Chen, Z. Parekh, H. Pham, ... and T. Duerig, “Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision,” in *Proceedings of the 38th International Conference on Machine Learning*, Online, pp. 4904-4916, July 2021. https://doi.org/10.48550/arXiv.2102.05918

[7] H. Liu, C. Li, Q. Wu, and Y. J. Lee, “Visual Instruction Tuning,” arXiv:2304.08485v1, April 2023. https://doi.org/10.48550/arXiv.2304.08485

[8] Z. Luo, G. Tang, C. Wang, Y. Zhou, X. Zheng, J. H. Wang, ... and D. Wu, “Generating High-quality Movie Tags from

Social Reviews: A Learning-driven Approach," in *Proceedings of 2021 IEEE International Conferences on Internet of Things (iThings) and IEEE Green Computing & Communications (GreenCom) and IEEE Cyber, Physical & Social Computing (CPSCom) and IEEE Smart Data (SmartData) and IEEE Congress on Cybermatics (Cybermatics)*, Melbourne, Australia, pp. 182-189, December 2021. https://doi.org/10.1109/iThings-GreenCom-CPSCom-SmartData-Cybermatics53846.2021.00040

[9] H. Park, S. Yong, Y. You, S. Lee, and I.-Y. Moon, "Automatic Movie Tag Generation System for Improving the Recommendation System," *Applied Sciences*, Vol. 12, No. 21, 10777, November 2022. https://doi.org/10.3390/app122110777

[10] S. Kar, S. Maharjan, A. P. López-Monroy, and T. Solorio, "MPST: A Corpus of Movie Plot Synopses with Tags," in *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, pp. 1734-1741, May 2018. https://doi.org/10.48550/arXiv.1802.07858

[11] J. Giles, "Internet Encyclopaedias Go Head to Head," *Nature*, Vol. 438, pp. 900-901, December 2005. https://doi.org/10.1038/438900a

[12] C. L. Weible, "The Internet Movie Database," *Internet Reference Services Quarterly*, Vol. 6, No. 2, pp. 47-50, 2001. https://doi.org/10.1300/J136v06n02_05

[13] J. Vig, S. Sen, and J. Riedl, "The Tag Genome: Encoding Community Knowledge to Support Novel Interaction," *ACM Transactions on Interactive Intelligent Systems (TiiS)*, Vol. 2, No. 3, 13, September 2012. https://doi.org/10.1145/2362394.2362395

[14] E. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, ... and W. Chen, "LoRA: Low-Rank Adaptation of Large Language Models," arXiv:2106.09685, 2021. https://doi.org/10.48550/arXiv.2106.09685

김현민(Hyun-min Kim)

2025년 : 가천대학교 AI·소프트웨어학부 (학사)

2019년～현 재: 가천대학교 AI·소프트웨어학부 학사과정

※관심분야 : 자연어처리(Natural Language Processing) 등

유준(Joon Yoo)

1997년 : KAIST 기계공학과 공학사

2009년 : 서울대학교 전기컴퓨터공학부 공학박사

2009년～2010년: University of California, Los Angeles (UCLA) 박사후 연구원

2010년～2012년: Bell Labs Seoul, Nokia, 책임 연구원

2012년～현 재: 가천대학교 AI·소프트웨어학부, 교수.

※관심분야 : 무선차량통신망(Vehicular Networks), 와이파이(Wi-Fi), 자연어처리(Natural Language Processing) 등