멀티모달 융합 기반 영화 장르 예측 모델 개발 및 성능 평가: MM-IMDb 데이터셋을 활용한 연구

Multi-modal Movie Genre Prediction Model Development and Performance Evaluation using the MM-IMDb Dataset

2025년 12월

서강대학교 AI∙SW대학원

인공지능 전공

민선기

차 례

1. 서론
   1. 연구 배경 및 동기
   2. 연구 목표
   3. 논문 구성
2. 관련 연구 및 배경
   1. 선행연구
   2. 멀티모달 영화 장르 예측과 MM-IMDb 데이터셋
   3. 멀티모달 융합 전략
   4. Vision Transformer (ViT)
   5. BERT 및 RoBERTa
   6. 객체 탐지 모델 (YOLO, Faster R-CNN)
3. 연구방법론
   1. 연구 개요
   2. 데이터셋
   3. 데이터 전처리
   4. 모달리티별 특징 추출
   5. 멀티모달 융합 전략
   6. 학습 및 평가 설정
   7. 설명가능성 (XAI) 적용
4. 실험 (TBD)
5. 결과 및 분석 (TBD)

참고문헌

초록

본 연구는 영화 포스터 이미지와 줄거리 텍스트를 통합적으로 활용하여 영화 장르를 예측하는 멀티모달 인공지능 모델을 제안한다. 기존 영화 장르 분류 연구는 주로 텍스트 또는 이미지 단일 모달리티에 의존하거나 단순한 Early Fusion, Late Fusion을 적용해왔다. 그러나 이러한 접근은 모달 간의 상호작용 정보를 충분히 반영하지 못한다는 한계를 가진다. 본 연구에서는 Cross-Attention 기반 멀티모달 Fusion 구조를 제안하여 포스터 이미지와 줄거리 텍스트 간의 복합적 연관성을 학습한다. 또한 baseline 비교를 위해 CNN, ViT, BERT와 같은 기존 모델과 더불어 YOLO, Faster R-CNN 기반 객체 감지 특징 추출 모델을 추가하여 성능을 검증한다. 제안 모델은 MM-IMDb 데이터셋의 23개 멀티라벨 장르 분류 문제를 대상으로 실험되며, 성능 비교를 통해 Cross-Attention Fusion의 효용성과 설명 가능성을 함께 확보하고자 한다.

1. 서론
   1. 연구 배경 및 동기

영화 장르는 콘텐츠 추천과 분류의 기본 단위로, 온라인 서비스, 영화 검색 플랫폼, 추천 시스템 등에서 중요한 역할을 한다. 기존 영화 장르 분류 연구는 주로 텍스트 기반(줄거리) 자연어 처리 또는 포스터 이미지 기반 컴퓨터 비전에 초점을 맞추어왔다. 그러나 영화 장르는 내러티브와 시각적 스타일의 결합으로 형성되는 만큼, 두 가지 모달리티를 동시에 활용하는 멀티모달 접근이 필요하다.

최근 Transformer 기반 모델(BERT, ViT)과 멀티모달 융합 전략(Attention, CLIP 등)이 활발히 연구되며 성능이 크게 향상되었다. 하지만 영화 장르 예측에서 이미지와 텍스트 간 복잡한 상호관계를 충분히 반영한 연구는 여전히 부족하다. 이에 본 연구는 Cross-Attention 융합 전략을 활용하여 장르 예측 성능과 해석 가능성을 동시에 향상시키고자 한다.

* 1. 연구 목표

본 연구의 최종 목표는 영화 포스터 이미지와 줄거리 텍스트를 동시에 활용하여 영화의 장르를 자동으로 분류하는 멀티모달 모델을 개발하는 것이다. 이를 위해 첫째, MM-IMDb 데이터셋에 포함된 이미지와 텍스트 데이터를 학습에 적합한 형태로 전처리하고, 둘째, 각 모달리티에서 최신 딥러닝 기반 인코더를 사용하여 고차원 표현을 추출한다. 셋째, Early Fusion, Late Fusion, Attention Fusion 등 기존 방식들과 제안하는 Cross-Attention Fusion 구조를 동일한 조건에서 비교 검증한다. 넷째, YOLO와 Faster R-CNN에서 탐지된 포스터 내 객체 특징을 baseline으로 도입하여 추가 비교를 수행함으로써 이미지의 세부 요소가 장르 예측에 어떤 기여를 할 수 있는지 확인한다. 마지막으로, Grad-CAM과 Attention Map 기반 시각화를 통해 모델 예측 근거를 분석하여 설명 가능성 확보까지 지향한다.

* 1. 논문 구성

본 논문은 총 여섯 장으로 구성된다.

제1장은 연구의 동기와 필요성, 연구 목표를 기술한다.

제2장은 멀티모달 영화 장르 분류와 관련된 기존 연구를 종합적으로 검토하고, 본 연구에서 적용하는 핵심 기술들의 배경을 상세히 설명한다.

제3장은 연구 방법론을 다루며, 데이터셋, 전처리 방식, 인코더 구조, 융합 전략, 학습 환경 등을 구체적으로 기술한다.

제4장은 제안 모델을 검증하기 위한 실험 설계와 비교 모델군을 소개한다.

제5장은 실험 결과를 정량적·정성적으로 분석하고, 모델의 해석 가능성에 대해 논의한다.

마지막으로 제6장은 본 연구를 요약하고 주요 기여를 정리하며, 향후 연구 방향을 제안한다.

1. 관련 연구 및 배경

2.1. 선행연구

멀티모달 학습(multimodal learning)은 서로 다른 유형의 데이터를 결합하여 더 풍부한 의미를 학습하는 인공지능의 핵심 연구 영역으로 자리 잡아왔다. 기존의 연구들은 텍스트와 이미지를 각각 독립적으로 처리하거나 단순히 결합하는 수준에 머물렀으나, 실제 사회적 맥락에서는 언어와 시각 정보가 동시에 존재한다는 점에서 두 모달리티를 결합한 학습이 점차 필요해졌다.

대표적인 사례로 Radford 등(2021)이 제안한 CLIP은 대규모 이미지-텍스트 쌍 데이터를 학습해 서로 다른 모달리티를 의미 공간에서 정렬하는 데 성공하였다. CLIP은 새로운 이미지와 설명 텍스트에 대해 사전 학습을 통해 획득한 지식을 활용하여 추가 학습 없이도 제로샷(Zero-shot) 분류를 수행할 수 있다는 점에서 큰 주목을 받았다. 이어서 Jia 등(2021)은 ALIGN 모델을 제시하여 웹에서 수집한 노이즈가 많은 대규모 이미지-텍스트 데이터를 통해 보다 일반화된 멀티모달 표현 학습을 가능케 했다. 또한 Li 등(2022)은 BLIP 모델을 통해 이미지 설명 생성(Image Captioning)과 시각적 질문 응답(VQA) 등 다중 과제에서 활용할 수 있는 통합형 사전학습 구조를 제안하였고, Liu 등(2023)의 LLaVA는 대규모 언어 모델(LLM)에 시각적 이해 능력을 결합하여 멀티모달 대화형 AI까지 확장하는 성과를 보여주었다.

이와 같은 최근 연구들은 단순히 텍스트와 이미지를 합치는 수준을 넘어서, 두 모달 간에 정교한 상호작용을 학습하는 방향으로 진화하고 있다. 특히 Kiela 등(2019)의 연구에서 논의된 바와 같이, Transformer 기반의 Cross-Attention 구조는 텍스트 내 단어 토큰과 이미지 내 특정 영역이 직접 연결되어 의미를 교차적으로 학습할 수 있게 한다. 이러한 흐름은 광고 분석, 의료 영상과 보고서 결합, 소셜미디어 멀티모달 분석 등 다양한 실질적 응용으로 확장되며 멀티모달 연구의 핵심적 발전을 이끌고 있다.

2.2. 멀티모달 영화 장르 예측과 MM-IMDb 데이터셋

MM-IMDb는 영화 장르 예측을 위한 대표적인 멀티모달 데이터셋으로, 약 25,000편 이상의 영화를 포함하고 있으며 각 영화에는 포스터 이미지, 줄거리 텍스트, 그리고 23개의 장르 멀티라벨이 제공된다 (Arevalo et al., 2017; Innovatiana, 2025). 이 데이터셋은 한 영화가 복수의 장르에 동시에 속할 수 있도록 구성되어 있으며, 이는 실제 영화 산업의 속성을 잘 반영한다. 기존 연구들은 CNN 및 ResNet 기반 모델을 활용하여 포스터 이미지를 인코딩하고, Word2Vec, GloVe, 그리고 최근에는 BERT와 같은 Transformer 기반 모델을 사용하여 줄거리 텍스트를 임베딩한 뒤 단순 결합하는 방식을 채택하는 경우가 많았다. 그러나 이러한 방식은 이미지와 텍스트 사이의 깊이 있는 상호작용을 충분히 반영하지 못하는 한계가 존재한다. 따라서 본 연구에서는 MM-IMDb 데이터셋을 바탕으로 이러한 약점을 극복하고자 하며, 특히 Cross-Attention 구조를 적용하여 기존 방식보다 더 풍부한 모달 간 연결 구조를 구현하고자 한다.

2.3. 멀티모달 융합 전략

멀티모달 학습의 중심 과제는 서로 다른 모달리티로부터 추출된 특징들을 효과적으로 결합하는 것이다. 기존 연구에서 가장 단순한 접근은 Early Fusion 방식이었다(Li et al., 2020). Early Fusion은 텍스트와 이미지에서 얻은 특징 벡터를 단순히 이어 붙여 하나의 입력으로 분류기에 전달하는 방식이다. 구조가 단순하고 계산량이 적다는 장점이 있으나, 모달 간 복잡한 의미적 관계를 충분히 반영하지 못한다는 한계를 가진다.

Late Fusion은 각 모달리티를 독립적으로 학습시킨 후 최종 단계에서 예측 확률이나 결과를 병합하는 방식이다(Kiela et al., 2019). 장점은 특정 모달에서 잡음이나 결손이 발생하더라도 다른 모달의 학습 결과가 이를 보완할 수 있다는 점이다. 그러나 두 모달 간의 상관관계를 직접적으로 학습하지 않기 때문에 융합의 효과가 제한적이라는 문제가 있다.

이후 Attention 메커니즘을 활용한 ****Attention Fusion****이 제안되었다. Attention Fusion은 모달리티별 특징 가운데 중요한 요소에 가중치를 두어 학습의 효율성을 높인다. 예를 들어 텍스트와 이미지를 동시에 사용하여 감정 분류를 수행하는 상황에서, 텍스트가 내용의 정서를 더 잘 드러내는 경우에는 텍스트에 높은 가중치가 할당되고, 반대로 이미지가 감정 신호를 더 강하게 반영하는 경우에는 이미지에 높은 가중치가 배분된다. 이러한 방식은 기존 단순 병합보다 정교하지만, 두 모달리티가 구조적으로 깊은 상호작용을 수행하지는 못한다는 한계가 지적된다.

최근에는 Transformer 구조를 확장한 ****Cross-Attention**** 방식이 널리 활용되고 있다(Kiela et al., 2019; Li et al., 2020). Cross-Attention에서는 텍스트와 이미지가 각각 Query, Key, Value 역할을 수행하면서 서로 정보를 교환한다. 즉, 텍스트 토큰이 이미지의 특정 패치와 상호작용하거나, 이미지의 세부 영역이 텍스트의 특정 단어와 대응하여 의미를 공유하게 된다. 이런 방식은 단순한 가중치 조정이 아니라, 두 모달리티 사이에 심층적인 교차 의미 학습을 가능케 하므로 최근 멀티모달 분류, 검색, 생성 연구에서 가장 각광받는 융합 전략 중 하나로 자리 잡았다.

2.4. Vision Transformer(ViT)와 ResNet50

이미지 처리 분야는 전통적으로 합성곱 신경망(CNN, Convolutional Neural Network)을 기반으로 발전해왔다. ****ResNet50****은 CNN 아키텍처에서 가장 대표적인 모델 중 하나로, ‘잔차 연결(residual connection)’을 도입하여 매우 깊은 네트워크에서도 기울기 소실 문제를 해결한 구조이다(He et al., 2016). 이 모델은 국소적인 이미지 특징을 안정적으로 추출할 수 있으며, 특정 객체, 질감, 패턴과 같은 부분적 단서 인식에 강점을 가진다. 예를 들어 영화 포스터에서 무기, 특정 의상 스타일, 로고 같은 요소는 지역적인 특징이기 때문에 ResNet50으로 잘 인식될 수 있다.

반면 최근 제안된 Vision Transformer(ViT)는 Transformer 구조를 시각적 데이터에 적용한 모델이다(Dosovitskiy et al., 2021). ViT는 입력 이미지를 작은 패치 단위로 분할하여 각 패치를 임베딩 벡터로 변환하고, 이를 텍스트 처리와 유사하게 Self-Attention 구조에 입력한다. 이 과정에서 이미지 전체의 전역적 문맥을 효과적으로 학습할 수 있다. 이는 CNN이 주로 지역적인 패턴 인식에 강점을 가진 반면, ViT는 이미지 전체의 배치, 색감, 레이아웃과 같은 전역적 관계를 포착하는 데 강점을 갖는다. 따라서 두 모델은 상호 보완적인 관계에 있으며, 본 연구에서는 ViT와 ResNet50을 모두 활용하여 이미지 모듈의 성능을 비교하고 각 모델이 제공하는 특징적 차이를 분석한다.

2.5. BERT 및 RoBERTa

자연어 처리(NLP) 분야에서는 Transformer 기반 언어 모델의 등장이 혁신적 변화를 이끌었다. 그중 \*\*BERT(Bidirectional Encoder Representations from Transformers)\*\*는 문맥을 양방향으로 학습하는 구조를 통해 단어의 의미를 정밀하게 추출할 수 있도록 고안되었다(Devlin et al., 2019). 예를 들어 “dark”라는 단어가 특정 문맥에서 ‘조도가 낮다’는 의미일 수도 있고, 다른 문맥에서 ‘음산하다’라는 정서를 드러낼 수도 있는데, BERT는 앞뒤 문맥 정보를 모두 고려하여 이런 의미 차이를 구체적으로 반영할 수 있다.

이후 BERT의 개선형 모델로 \*\*RoBERTa(Robustly Optimized BERT Pretraining Approach)\*\*가 제안되었다(Liu et al., 2019). RoBERTa는 더 많은 학습 데이터와 더 긴 훈련 기간을 도입하고, 사전 학습 시 마스킹 전략을 개선하여 BERT보다 표현력이 강력한 모델로 발전했다. 그 결과, 다양한 자연어 처리 과제에서 일관되게 BERT보다 높은 성능을 달성했다. 이러한 특성으로 인해 BERT와 RoBERTa는 단순 분류를 넘어, 줄거리 텍스트와 같은 서사적 자료에서 등장인물, 사건, 정서적 분위기 등 장르를 암시하는 다양한 단서를 효과적으로 추출할 수 있는 핵심 모델로 활용되고 있다.

2.6. 객체 탐지 모델 (YOLO, Faster R-CNN)

이미지에서 의미 있는 객체를 식별하는 작업은 시각적 이해의 필수적인 요소이며, 멀티모달 학습에서도 중요한 역할을 한다. \*\*YOLO(You Only Look Once)\*\*는 단일 단계(one-stage) 탐지 방식으로 이미지를 한 번의 연산 과정에서 처리하여 매우 빠른 객체 탐지를 가능케 한다(Redmon et al., 2016). 이러한 특성은 실시간 객체 인식이 필요한 환경에 적합하며, 영화 포스터와 같이 다양한 시각적 요소가 혼합된 이미지에서도 인물이나 특정 물체와 같은 주요 객체를 빠르게 포착할 수 있다.

반면 ****Faster R-CNN****은 두 단계(two-stage) 탐지 방식을 채택하여 더 정밀한 인식을 가능하게 한다(Ren et al., 2015). 첫 번째 단계에서는 영역 제안 네트워크(Region Proposal Network)가 후보 영역을 탐색하고, 두 번째 단계에서는 이 후보 영역을 정밀하게 분류한다. 이러한 구조는 탐지 속도는 YOLO보다 느리지만 작은 객체와 복잡한 배경 속 물체까지 높은 정확도로 탐지할 수 있다는 장점이 있다. 영화 포스터 분석에서는 무기, 배경의 상징적 사물, 특정 인물 등 장르적 단서를 나타내는 다양한 객체가 존재하기 때문에 이러한 탐지 모델은 단순 시각적 특징 추출 이상의 의의를 갖는다. 결과적으로 YOLO와 Faster R-CNN은 포스터 이미지에 담긴 의미적 요소를 보조적으로 추출하여 텍스트 정보와 결합함으로써 멀티모달 학습의 성능을 강화할 수 있다.

1. 연구방법론
   1. 연구 개요

본 연구의 방법론은 크게 네 단계로 구성된다. 먼저 데이터 전처리를 통해 MM-IMDb 데이터셋의 이미지와 텍스트를 분석 가능한 형태로 가공한다. 다음으로 각 모달리티에서 대표적인 딥러닝 모델들을 활용하여 특징을 추출한다. 이후 멀티모달 융합 단계에서는 다양한 융합 전략을 비교하고, 제안하는 Cross-Attention 기반 구조를 설계한다. 마지막으로 모델 학습 및 평가를 수행하면서 성능을 검증하고, 설명 가능한 AI(XAI) 기법을 통해 모델의 해석 가능성까지 확보한다.

* 1. 데이터셋

연구에 사용되는 데이터셋은 MM-IMDb로, 약 25,000편의 영화를 포함하고 있으며 포스터 이미지와 줄거리 텍스트, 그리고 23개의 멀티라벨 장르 정보를 제공한다. 각 샘플은 다중 장르 분류 문제를 구성할 수 있는 구조를 갖추고 있어 본 연구의 목적에 적합하다. 데이터셋은 학습용, 검증용, 테스트용으로 각각 70%, 15%, 15%의 비율로 분할하여 모델 훈련과 성능 평가에 활용된다. 이러한 구성은 데이터셋의 균형성을 보장하면서 과적합을 방지하기 위한 것이다.

* 1. 데이터 전처리

포스터 이미지는 학습 효율성을 높이기 위해 224×224의 크기로 리사이즈하고, 픽셀값을 정규화한다. 데이터의 일반화 성능을 높이기 위해 랜덤 크롭, 좌우 반전, 색상 변화와 같은 증강 기법도 일부 적용한다. 줄거리 텍스트는 소문자 표준화, 불용어 및 특수문자 제거 과정을 거친 뒤 BERT 토크나이저에서 요구하는 형식으로 변환된다. 각 문장은 최대 512 토큰 길이로 제한되며, 길이가 부족한 경우 padding이 적용된다. 이러한 전처리 과정을 통해 데이터는 학습 모델에 최적화된 형태로 정제된다.

* 1. 모달리티별 특징 추출

이미지 모달리티에서는 CNN 계열의 ResNet50과 Transformer 계열의 Vision Transformer(ViT)를 활용한다. ResNet은 이미지의 지역적 특징을 안정적으로 포착할 수 있으며, ViT는 포스터 전체의 전역적 문맥을 반영하는 데 효과적이다. 텍스트 모달리티에서는 BERT와 RoBERTa를 사용한다. 두 모델은 Transformer 기반 언어모델로, 영화 줄거리와 같은 긴 서사적 텍스트에서 핵심적 의미를 효과적으로 학습할 수 있다는 장점을 가진다. 또한 baseline 모델군에는 YOLO와 Faster R-CNN을 포함시켜, 포스터 내 주요 객체(예: 무기, 인물, 배경 요소)를 탐지하고 이를 벡터 형태의 부가 특징으로 변환한다. 이 과정을 통해 추출된 객체 정보는 장르 분류의 보조 단서로 사용된다.

* 1. 멀티모달 융합 전략

본 연구에서는 다양한 멀티모달 융합 전략을 구현하여 성능을 비교한다. 먼저 Early Fusion은 이미지와 텍스트 특징 벡터를 단순히 병합하여 분류기에 입력하는 방식이다. Late Fusion은 독립적으로 학습된 각 모달리티 분류 결과를 결합하여 최종 예측으로 반영한다. Attention Fusion은 모달리티별 중요도를 가중치로 부여하여 두 입력의 결합을 보다 유연하게 조정한다. 마지막으로 본 연구에서 제안하는 Cross-Attention Fusion은 텍스트와 이미지가 서로 Query, Key, Value로 작용하며 상호 정보를 교환하는 구조이다. 이를 통해 포스터에서 추출된 시각적 단서와 줄거리 텍스트 속 의미적 단서를 함께 학습하여, 기존 방법들보다 정교한 장르 예측이 가능하게 한다.

* 1. 학습 및 평가 설정

모델 학습은 PyTorch 프레임워크에서 수행되며 GPU 기반 연산을 활용한다. Optimizer로는 AdamW가 사용되며, 학습 안정성을 위해 learning rate warm-up을 적용한다. 손실 함수는 멀티라벨 분류에 적합한 Binary Cross-Entropy with Logits Loss를 활용한다. 평가 지표로는 전체 정확도(Accuracy), 장르별 정밀도(Precision), 재현율(Recall), F1-score, ROC-AUC, mean Average Precision(mAP)을 사용하여 성능을 다각도로 측정한다.

* 1. 설명가능성 (XAI) 적용

모델의 해석 가능성을 확보하기 위해 Grad-CAM을 활용하여 포스터 이미지 중 어느 영역이 장르 예측에 기여했는지를 시각적으로 강조한다. 또한 Cross-Attention 융합 구조에서 산출된 Attention 가중치를 분석함으로써, 줄거리 텍스트의 어떤 단어가 특정 장르 예측에 결정적 역할을 했는지 보여준다. 이를 통해 단순 성능 비교를 넘어, 모델이 실제로 어떤 요소를 근거로 판단하는지 설명할 수 있으며, 이는 실질적 응용 및 연구 기여도 측면에서 중요한 의미를 가진다.

1. 실험 (TBD)

4.1 실험 목적

본 연구의 실험 목적은 제안된 Cross-Attention 기반 멀티모달 장르 예측 모델이 기존의 융합 방식과 비교했을 때 실제로 더 높은 성능과 해석 가능성을 제공하는지 검증하는 데 있다. 기존의 Early Fusion은 단순한 결합 방식이었고, Late Fusion은 모달리티 독립성을 강조했으며, Attention Fusion은 가중치를 활용하여 유연성을 부여했지만 여전히 모달 간 풍부한 상호작용은 학습하기 어려웠다. 또한 이미지 단독 기반 모델(CNN, ViT), 텍스트 단독 기반 모델(BERT), 객체 탐지 기반 모델(YOLO, Faster R-CNN) 역시 부분적인 정보에만 의존하기 때문에 장르 예측에서 한계를 보인다. 따라서 본 실험에서는 이러한 다양한 baseline 모델과 비교하여 Cross-Attention Fusion 구조의 우수성을 입증하는 것을 목표로 한다.

4.2 비교 모델군

본 연구의 실험에 포함되는 비교 모델군은 크게 다섯 가지 범주로 나눌 수 있다.

첫째, 텍스트 단일 모달리티 기반 모델로 BERT와 RoBERTa를 활용하여 줄거리 서사를 독립적으로 학습한 분류 모델을 설정하였다. 이는 기존 연구에서 가장 보편적으로 사용된 접근 방식으로, 본 연구의 성능 비교 기준점이 된다.

둘째, 이미지 단일 모달리티 기반 모델로는 CNN 계열의 ResNet50과 Transformer 기반의 Vision Transformer(ViT)를 사용한다. ResNet은 포스터의 지역적 특징을 잘 포착하는 장점이 있고, ViT는 이미지의 전역적 문맥 정보를 학습하는 능력이 탁월하다.

셋째, 객체 탐지 기반 모델을 baseline으로 포함한다. YOLO는 빠르고 효율적인 one-stage 탐지 모델이며, Faster R-CNN은 두 단계 탐지를 통해 높은 정확도를 보이는 모델이다. 두 모델을 활용하여 영화 포스터에서 인물, 무기, 배경과 같은 중요한 객체를 추출하고 이를 특징 벡터로 변환하여 장르 분류 성능에 반영함으로써, 객체 단서의 기여도를 확인한다.

넷째, 기존 멀티모달 융합 모델로 Early Fusion, Late Fusion, Attention Fusion을 구현한다. 이는 각각의 융합 전략이 Cross-Attention과 비교했을 때 어떤 성능 차이를 보이는지 검증하기 위해 사용된다.

마지막으로, 제안 모델인 Cross-Attention 기반 Fusion 구조를 학습시킨다. 이 모델에서는 줄거리의 텍스트 토큰과 포스터의 이미지 패치가 서로 Attention을 주고받으며 상호작용을 학습한다. 이를 통해 기존 모델보다 더 정교한 멀티모달 의미 관계를 파악할 수 있을 것으로 기대된다.

4.3 실험 절차

본 연구의 실험 절차는 다음과 같은 일련의 과정으로 이루어진다.

첫째, 데이터 로딩 및 전처리 단계에서는 MM-IMDb 데이터셋을 불러오고, 이미지 전처리(리사이징, 정규화, 데이터 증강)와 텍스트 전처리(BERT 토크나이징, padding, truncation)를 수행한다.

둘째, 추출 단계에서는 이미지 모달리티에서 ResNet50, ViT, YOLO, Faster R-CNN을 통해 시각적 특징을 도출하고, 텍스트 모달리티에서는 BERT와 RoBERTa를 통해 언어 표현 벡터를 추출한다.

셋째, 융합 단계에서는 Early Fusion, Late Fusion, Attention Fusion, Cross-Attention Fusion 네 가지 방식을 각각 적용한다. 특히 Cross-Attention에서는 Query-Key-Value Attention 메커니즘을 통해 텍스트와 이미지가 서로 교차적으로 정보를 주고받으며 결합된다.

넷째, 분류 단계에서는 결합된 멀티모달 벡터를 Fully Connected Layer 기반 분류기에 입력하여 23개 영화 장르에 대한 멀티라벨 분류를 수행한다. 이때 Binary Cross-Entropy with Logits Loss를 사용하여 최적화한다.

다섯째, 평가 단계에서는 Accuracy, Precision, Recall, F1-score, ROC-AUC, mAP와 같은 다양한 지표를 활용하여 모델 성능을 정량적으로 측정한다. 또한 Grad-CAM과 Attention Map을 통해 포스터의 특정 영역이나 줄거리의 특정 단어가 예측에 영향을 미친 정도를 시각화하여 정성적 분석을 병행한다.

이 과정을 통해 본 연구에서는 다양한 baseline들과 비교하여 Cross-Attention Fusion 모델의 상대적 성능 차이를 명확하게 도출할 수 있으며, 나아가 모델이 학습한 멀티모달 의미 관계의 타당성을 설명할 수 있다.

1. 결과 및 분석 (TBD)

참고문헌

Radford, A. et al. (2021). Learning Transferable Visual Models From Natural Language Supervision (CLIP). ICML. [**https://arxiv.org/abs/2103.00020**](https://arxiv.org/abs/2103.00020)

Jia, C. et al. (2021). Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision (ALIGN). ICML. [**https://arxiv.org/abs/2102.05918**](https://arxiv.org/abs/2102.05918)

Li, J. et al. (2022). BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. ICML. [**https://arxiv.org/abs/2201.12086**](https://arxiv.org/abs/2201.12086)

Liu, H. et al. (2023). Visual Instruction Tuning for LLMs (LLaVA). [**https://arxiv.org/abs/2304.08485**](https://arxiv.org/abs/2304.08485)

Kiela, D. et al. (2019). Supervised Multimodal Bitransformers for Classifying Images and Text. arXiv. [**https://arxiv.org/abs/1909.02950**](https://arxiv.org/abs/1909.02950)

Li, X. et al. (2020). Hierarchical multimodal fusion for movie genre classification. Information Fusion.

Shaip (2025). Multimodal AI – The Complete Guide to Training Data. [**https://ko.shaip.com/blog/multimodal-ai-the-complete-guide-to-training-data/**](https://ko.shaip.com/blog/multimodal-ai-the-complete-guide-to-training-data/)

He, K. et al. (2016). Deep Residual Learning for Image Recognition (ResNet). CVPR. [**https://arxiv.org/abs/1512.03385**](https://arxiv.org/abs/1512.03385)

Dosovitskiy, A. et al. (2021). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale (ViT). ICLR. [**https://arxiv.org/abs/2010.11929**](https://arxiv.org/abs/2010.11929)

Devlin, J. et al. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. NAACL. [**https://arxiv.org/abs/1810.04805**](https://arxiv.org/abs/1810.04805)

Liu, Y. et al. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv. [**https://arxiv.org/abs/1907.11692**](https://arxiv.org/abs/1907.11692)

Redmon, J. et al. (2016). You Only Look Once: Unified, Real-Time Object Detection. CVPR. [**https://arxiv.org/abs/1506.02640**](https://arxiv.org/abs/1506.02640)

Ren, S. et al. (2015). Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. NeurIPS. [**https://arxiv.org/abs/1506.01497**](https://arxiv.org/abs/1506.01497)