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

영화 장르 예측에 대한 초기 연구들은 주로 텍스트 기반의 자연어 처리 기법에 의존하였다. 영화 줄거리 텍스트를 활용한 장르 분석은 문서 분류 접근법과 RNN이나 CNN 같은 신경망 기법을 중심으로 이루어졌다. 그러나 영화 장르를 결정짓는 데에는 시각적 단서 또한 중요한 역할을 한다는 점이 강조되면서 이미지와 텍스트를 함께 고려하는 멀티모달 연구가 등장하게 되었다. Arevalo 등(2017)은 이를 위해 MM-IMDb 데이터셋을 제안하며 CNN 기반 이미지 인코더와 RNN 기반 텍스트 인코더를 Early Fusion으로 결합해 장르를 분류하였다 (Arevalo et al., 2017). 이후 Li 등(2020)은 Late Fusion과 Attention 기반 융합 방식을 도입하여 모달 단위 예측 결과를 결합함으로써 성능을 향상시켰고 (Li et al., 2020), Kiela 등(2019)은 Attention 메커니즘을 포함한 멀티모달 Transformer 구조를 제안하며 텍스트와 이미지 간 연관성을 더 효과적으로 학습할 수 있음을 보여주었다 (Kiela et al., 2019). 최근에는 CLIP이나 LLaVA 등 대규모 Vision-Language 모델(VLM)이 영화 장르 예측에 적용될 가능성을 보이고 있으나, MM-IMDb 데이터셋을 대상으로 Cross-Attention 융합 전략을 본격적으로 활용한 연구는 아직 드문 상황이다.

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 학습에서 가장 큰 도전은 서로 다른 모달리티로부터 추출된 특징을 효과적으로 결합하는 것이다. 기존에 널리 연구된 융합 전략은 Early Fusion, Late Fusion, Attention Fusion으로 구분할 수 있다 (Li et al., 2020; Kiela et al., 2019). Early Fusion은 텍스트와 이미지의 임베딩 벡터를 단순히 하나의 벡터로 병합하는 방법으로 구현이 간단하다는 장점이 있으나, 모달 간 관계를 충분히 반영하지 못한다. Late Fusion은 텍스트와 이미지 각각을 독립적으로 학습시킨 다음, 각 분류 결과를 결합하는 방식으로 모달 독립성을 강조할 수 있지만 모달 간 상호작용 정보는 제한된다. Attention Fusion은 모달리티별로 중요도를 계산해 가중치를 반영하는 방식으로, 단순 병합이나 독립 결합보다는 정교하지만 여전히 제한적이다. 이에 비해 Cross-Attention은 텍스트 토큰과 이미지 패치가 서로 Query, Key, Value 역할을 주고받으며 상호작용하기 때문에 모달 간 의미를 교차적으로 학습할 수 있다. 2025년 연구 동향에서도 멀티모달 AI 연구의 핵심은 데이터 융합이며 복수 모달 간 상호작용을 최적화하는 방식으로 발전하고 있음을 지적하고 있어 (Shaip, 2025; Web Search Result 1), Cross-Attention은 매우 유망한 방법론이라 할 수 있다.

2.4. Vision Transformer (ViT)

Vision Transformer(ViT)는 이미지를 일정한 크기의 패치로 나누어 각각을 토큰으로 취급하고, 이를 Transformer의 Self-Attention 메커니즘에 입력하여 이미지를 처리하는 구조이다. 이 모델은 CNN이 주로 지역적 특징을 잘 학습하는 반면, 전역적인 문맥 정보를 포착하는 데에 강점을 가진다. 영화 포스터와 같은 시각 자료에는 색감, 인물 배치, 배경 처리와 같은 전역적 요소가 장르를 결정짓는 중요한 단서가 되므로, ViT는 영화 장르 분류에 적합한 선택이라고 할 수 있다 (Dosovitskiy et al., 2021).

2.5. BERT 및 RoBERTa

자연어 처리 영역에서는 Transformer 기반 언어모델이 영화 줄거리와 같은 문맥적 서사 텍스트를 처리하는 데 큰 성과를 거두었다. BERT는 양방향적으로 문맥을 학습하여 문장 내 단어의 의미를 보다 정밀하게 추출할 수 있으며 (Devlin et al., 2019), RoBERTa는 BERT의 학습 절차를 최적화하고 더 많은 데이터를 사용하여 개선된 성능을 보여준다 (Liu et al., 2019). 영화 줄거리에는 인물, 사건, 감정적 분위기 등 장르의 특성이 드러나는 중요한 신호들이 포함되어 있는데, BERT 계열 모델은 이러한 서사적 단서를 효과적으로 학습하여 장르 분류의 성능을 높일 수 있다.

2.6. 객체 탐지 모델 (YOLO, Faster R-CNN)

객체 탐지 모델은 이미지 속 특정 물체들을 인식하고 그 위치를 특정하는 기술로, 영화 포스터 분석에서도 유용하게 활용될 수 있다. YOLO(You Only Look Once)는 이미지를 한 번의 연산으로 처리하여 빠르고 효율적인 탐지가 가능하다는 점에서 강점을 가진다 (Redmon et al., 2016; Web Search Result 2). 반면 Faster R-CNN은 두 단계 접근 방식으로 후보 영역을 먼저 제안하고 이를 정밀하게 분류하기 때문에 속도는 느리지만 탐지 정확도가 높다는 장점을 가진다 (Ren et al., 2015). 영화 포스터는 특정 인물, 무기, 배경 요소 같은 시각적 단서가 장르 예측에 직접적인 영향을 미치므로, YOLO와 Faster R-CNN은 포스터에서 이러한 세부 요소를 자동으로 추출하여 baseline 모델로 유의미한 특징을 제공할 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

1. Arevalo, J., Solorio, T., Montes-y-Gómez, M., & González, F. (2017). *Gated Multimodal Units for Information Fusion*. ICLR.
2. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. NAACL.
3. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., … & Houlsby, N. (2021). *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*. ICLR.
4. Innovatiana (2025). *MM-IMDb Dataset: Multimodal Film Analysis*. Retrieved from <https://www.innovatiana.com/en/datasets/mm-imdb-multimodal-imdb-dataset>
5. Kiela, D., Bulgakov, V., & Cho, K. (2019). *Supervised Multimodal Bitransformers for Classifying Images and Text*. arXiv:1909.02950.
6. Li, X., Gong, Y., Li, J., & Peng, B. (2020). *Hierarchical multimodal fusion for movie genre classification*. *Information Fusion*.
7. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., … & Stoyanov, V. (2019). *RoBERTa: A Robustly Optimized BERT Pretraining Approach*. arXiv:1907.11692.
8. Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). *You Only Look Once: Unified, Real-Time Object Detection*. CVPR.
9. Ren, S., He, K., Girshick, R., & Sun, J. (2015). *Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks*. NeurIPS.
10. Shaip (2025). *Multimodal AI – The Complete Guide to Training Data*. Retrieved from [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)
11. Web Search Result 1: *Recent research in film genre prediction and object detection techniques summary* (2025).
12. Web Search Result 2: *Как обнаруживать объекты, используя YOLO , OpenCV и PyTorch...* (2021).