COSE474-2024F: Final Project
“영화 포스터에서 장르별 주목도 확인을 적용한 다중 분류 모델”

2022320009 이수현

1. Introduction

1.1. Motivation

영화 포스터는 영화의 특징과 분위기를 시각적으로 전달하며, 여러 장르를 동시에 반영한다. 본 연구는 장르가 여러 개인 한 영화의 포스터에 모델이 주목하는 영역이 장르별로 다를 것이라는 가정에서 출발한다. 이를 통해 영화 포스터가 가지는 다양한 시각적 특징을 분석하고, 장르별로 차별화된 특징을 효과적으로 해석하고자 한다.

1.2. Problem Definition

본 프로젝트는 영화 포스터 이미지를 입력으로 하여 다중 장르를 예측하는 멀티라벨 분류 문제를 다룬다. 단순히 다중 장르를 정확도 높게 예측하는 것이 아닌, 장르별로 포스터 이미지의 주목 위치의 차이를 파악하고 장르별로 어떻게 차이를 보이는지 분석한다. 본 연구의 주요 과제는 장르별 학습 과정을 통해 주목 지점의 차이를 파악하는 것과 각 장르의 특성을 반영한 다중 장르 예측 모델 개발하는 것이다.

1.3. Contributions

데이터셋 정제, 오버샘플링과 증강을 통한 장르별로 균형잡힌 데이터셋 확보 및 장르 임베딩을 활용한 Classifier Chain 기법을 적용하여 장르 간 상관관계를 반영한 멀티라벨 분류 모델 제안 한다. 또한 Attention 시각화를 통해 장르별 포스터 이미지의 주목 위치의 차이를 분석한다.

2. Proposed Method

2.1. Key Model Introduction: CLIP

본 연구에서는 CLIP (Guo, 2024)을 기본 모델로 사용한다. CLIP은 이미지와 텍스트를 공동 학습하여 복합적인 시각 정보를 학습할 수 있고, 대규모 데이터로 사전 학습된 모델로 강한 일반화 성능을 가지고 있다. 또한 CLIP인코더 (Ali & Khan, 2024) 백본 구조가 프로젝트에서 장르 간 시각적 차이 분석에 적합할 것이라고 기대했다.

2.2. Significance and Novelty

또한, 논문 (Agarwal et al., 2021)에서 강조된 바와 같이, 데이터 편향성을 줄이고 모델의 안정성을 높이기 위해 데이터 다양성과 균형을 중점적으로 고려하였다.

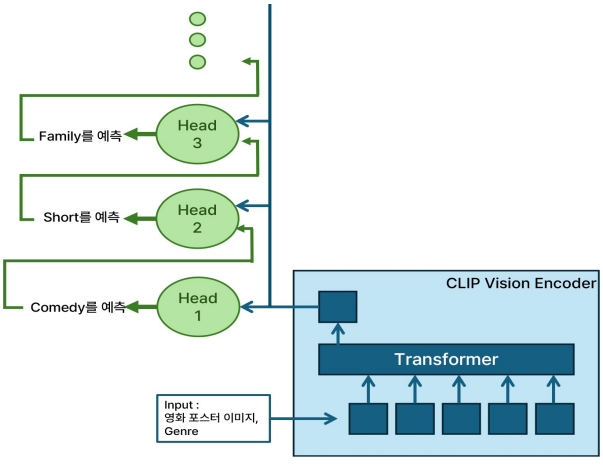
일반적인 멀티라벨 분류와 달리, 본 연구에서는 각 장르의 고유한 특징을 반영하기 위해 각 라벨을 독립적으로 학습하는 방식으로 진행하였다. 그러나 이런 방식으로 진행하면 각 장르 간 상관관계를 학습하지 못하기 때문에 Classifier Chain을 사용하였다. 라벨 순서가 성능에 영향을 미치기 때문에 CLIP의 텍스트 임베딩을 활용해 유사도가 높은 순서로 라벨을 정렬하여 관계를 효과적으로 학습하도록 설계하였다.

2.3. chanllenges

1. 데이터 불균형 문제 : 일부 장르는 10개 내외로 매우 적은 반면, 어떤 장르는 600개 이상으로 데이터 양의 불균형이 심각했다. 그래서 저빈도 장르를 제거하고, 데이터 증강과 오버샘플링을 통해 데이터를 많이 확보하고, 각 장르에서 250개씩 샘플링하여 균형을 맞추었다.

2. Classifier Chain의 라벨 순서 결정 : 최적의 라벨 순서를 위해 CLIP의 텍스트 임베딩을 활용하여 장르 간 의미적 유사도를 계산하고, 유사도가 높은 순서로 정렬한다.

2.4. Method Overview



모델은 영화 포스터 이미지를 입력받아 CLIP Vision Encoder에서 글로벌 특징 벡터를 생성한다. 특징벡터는 Head 1으로 전달되어 Comedy 장르를 예측하고, 이후 이전 Head의 예측값과 특징 벡터가 Head 2로 전달되어 Short 장르를 예측하는 방식으로 진행된다. 이 과정이 반복되며 Multi-head와 Classifier Chain 구조를 통해 각 장르를 순차적으로 학습한다.

2.5. Algorithm

2.5.1. Pseudocode

Algorithm 1 Multi-head with Classifier Chain Model

Input: Pixel values $\mathbf{X}$, labels $\mathbf{Y}$ (optional), vision model f, number of genres n
Initialize $chain\_input = f(\mathbf{X})$ (pooled output of vision model)
Initialize empty list $logits\_list$
for $i = 1$ **to** n **do**
  Compute logit: $logit = head_i(chain\_input)$
  Append $logit$ to $logits\_list$
  if labels $\mathbf{Y}$ are provided **then**
    $prev\_label\_info = \mathbf{Y}[:, i].unsqueeze(-1)$ (teacher forcing)
  else
    Compute prediction: $prev\_pred = (\sigma(logit) > 0.5)$
    $prev\_label\_info = prev\_pred$
  end if
  Update $chain\_input = concat(chain\_input, prev\_label\_info)$
end for
Concatenate logits: $\mathbf{L} = concat(logits\_list)$
Output: Logits $\mathbf{L}$

2.5.2. Features

- classifier chain으로 각 장르 간의 상관관계를 학습한다.

- 학습 단계에서는 정답 라벨을 이용하는 teacher forcing 기법으로 더 안정적인 학습을 지원한다.

- multihead구조로, 각 장르에 독립적인 로지스틱 회귀 헤드를 사용하여 각 장르의 특징을 개별적으로 학습한다.

2.5.3. Advantages and significance

- 개별장르를 학습에 라벨 간 관계 학습을 더하고, 멀티라벨 분류에 적합하다.

3. Experiments

3.1. Dataset

3.1.1. Dataset and purpose

본 연구에서는 Kaggle에서 제공되는 Movie Genre from its Poster을 사용하였다. 영화 포스터 이미지와 장르가 포함되어 있어 프로젝트의 목적에 적합한 데이터셋이다.

3.1.2. Dataset Composition and Characteristics

영화 포스터 이미지 (‘.jpg‘) 997개와 ID (‘imdbId‘), Genre 등의 정보가 있는 CSV 파일(‘MovieGenre.csv‘)로 구성되어 있으며 각 영화 포스터는 다중 장르를 가질 수 있다.

3.1.3. Data Preprocessing

프로젝트에 필요한 imdbId와 Genre열만 추출하여 사용하였고, 다중 장르를 각각의 장르로 분할하였다. 저빈도 장르는 제거하고, 데이터 증강 및 오버샘플링을 통해 데이터를 확보하였으며 데이터 균형을 위해 최종 데이터는 각 장르에서 250개씩 사용하였다. 입력 데이터로 들어갈 영화 포스터 이미지는 CLIP 모델의 크기에 맞게 224x224로 변환하였다.

3.1.4. Dataset Splitting

- 전체 데이터셋을 학습:검증:테스트 = 8:1:1로 분할하며 그 과정에서 각 세트에서 클래스 분포를 유지하기 위해 stratified sampling기법을 사용한다. - 분할 후 데이터 수는 대략 학습데이터(797개), 검증데이터(100개), 테스트데이터(100개)로 구성된다.

3.2. Computing Resources, Experimental design and Setup

1. 하드웨어 환경(구글 코랩 제공) - GPU : NVIDIA Tesla T4 / GPU 메모리: 15,360 MB / CPU : Intel Xeon 2.3GHz / RAM : 12GB

2. 소프트웨어 환경 - 운영체제: Ubuntu 22.04.3 LTS (Jammy Jellyfish) / PyTorch 버전: 2.5.1+cu121 / Torchvision 버전: 0.20.1+cu121 / CUDA Version: 12.2 / 기타 라이브러리: Pandas, NumPy, Matplotlib, Scikit-learn, Transformers

3. 데이터셋 - kaggle의 Movie Genre from its Poster

4. 하이퍼파라미터 - learning rate: 0.0005, oprimizer : Adam, batch size : 32, epoch : 10, loss function : BCEWithLogitsLoss

5. 평가지표 - precision, accuracy, F1-score, hamming loss, subset accuracy, Coverage Error

6. 실험 설계

데이터셋을 학습 8 : 검증 1 : 테스트 1 로 분할 / 계층 샘플링(stratified sampling)을 통해 클래스 분포를 유지 / Teacher Forcing 기법을 사용

7. 모델 아키텍처 - Classifier Chain + Multi-head

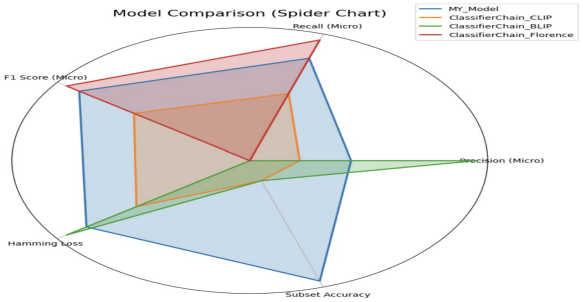
3.3. Quantitative Results

Precision은 양호한 수준을 보였으나, Recall은 부족한 결과로 나왔다. 멀티라벨 모델의 성능을 평가하는 지표를 위주로 분석해보자면, Hamming Loss는 낮은 값으로 전반적인 예측 성능은 양호하였으나, Subset Accuracy가 낮아 모든 라벨을 정확히 예측하는 데는 어려움이 있었음을 알 수 있다. 또한, Coverage Error는 평균적으로 15개 이상의 예측이 필요함을 보여주며, 모델이 일부 라벨을 놓치고 있음을 보여준다.

Metric	Type	Value
Precision	Micro	0.5349
	Macro	0.5891
Recall	Micro	0.3010
	Macro	0.2423
F1 Score	Micro	0.3853
	Macro	0.2911
Hamming Loss	-	0.1519
Subset Accuracy	-	0.0397
Coverage Error	-	15.0993

Table 1. My Model Evaluation Metrics

3.5. Analysis

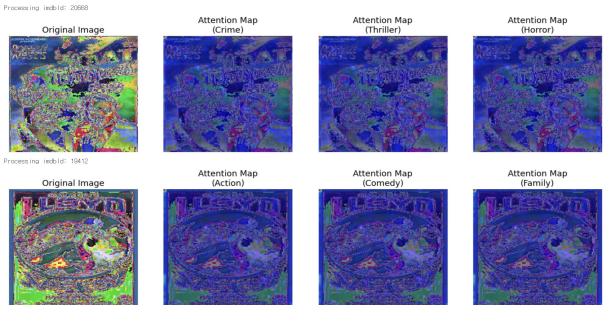


Model	Hamming Loss	Subset Accuracy	Coverage Error
MY_Model	0.157699	0.029801	15.503311
ClassifierChain_CLIP	0.160803	0.013245	15.768212
ClassifierChain_BLIP	0.156457	0.013245	15.807947
ClassifierChain_Florence	0.167839	0.009934	15.688742

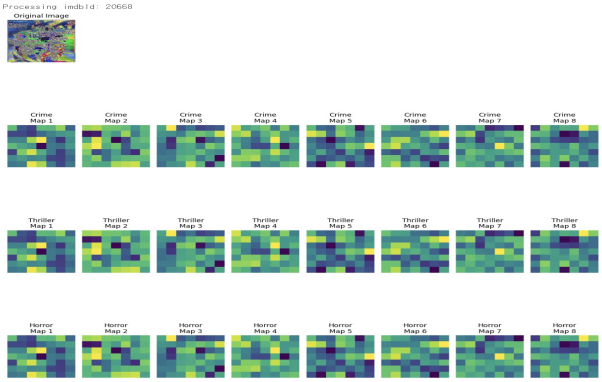
Table 2. Comparison of Hamming Loss, Subset Accuracy, and Coverage Error Across Models

다른 모델과 비교하여 본 프로젝트의 모델을 평가한다. F1 Score는 양호하고, coverage Error가 다른 모델들 보다 높은 15개 이상의 예측이 필요하며 이는 다중 장르 예측에서 모델이 세부 조정이 부족했음을 보여준다. subset accuracy가 가장 작은 값으로 다른 모델들에 비해 뛰어나지만, 절대적으로 작은 값으로 모든 라벨을 정확히 예측하기 어렵다는 한계를 드러낸다.

3.4. Qualitative Results



영화 포스터와 장르별 attention map 결과를 분석한 결과, 같은 영화 포스터에 대해 장르와 상관없이 모든 attention map이 유사하게 나타난다. 이는 모델이 장르와 관계없이 포스터의 동일한 요소에 집중하며, 각 장르를 구별할 수 있는 특징을 제대로 학습하지 못하였다는 것을 보여준다.



각 장르의 feature map에서 큰 차이가 발견되지 않고 넓은 영역에 걸쳐 비슷한 패턴을 보인다. 이는 모델이 특정 장르와 관련된 국소적인 특징보다는 전체적인 패턴에 의존하고 있음을 보여준다.

3.6. Discussion

프로젝트가 실패한 이유를 다음과 같이 분석할 수 있다.

1. 가정의 불확실성 : Attention Map과 Feature Map을 분석한 결과 모델이 장르별로 주목하는 포스터의 영역에 뚜렷한 차이가 나타나지 않는다. 장르별 시각적 특징 부족 또는 CLIP 인코더가 제대로 학습되지 못했을 수 있다.

2. 다중 장르의 복합성 : 상반되는 장르 조합의 포스터는 서로의 학습에 방해가 되었을 수 있다.

3. 과거 데이터의 한계 : 오래된 영화 포스터의 단순한 시각적 특징이 모델 학습에 어려움을 주었을 수 있다.

4. Future Directions

discussion에서 분석한 실패 원인을 토대로 연구를 발전시키기 위해 다음을 제안한다.

1. 더 적합한 모델 사용 : 장르별 고유한 시각적 특징을 잘 학습할 수 있는 모델과 GNN등 강한 장르 간 관계 학습 모델의 사용 및 더욱 정교한 attention 기법을 도입한다.

2. 영화 포스터 이미지와 함께 줄거리 요약을 입력 데이터로 넣어 포스터가 제공하지 못하는 부분을 보완한다.

3. 최신 영화 데이터를 추가하여 데이터셋을 확장한다.

References

Agarwal, S., Krueger, G., Clark, J., Radford, A., Kim, J., and Brundage, M. Evaluating clip: Towards characterization of broader capabilities and downstream implications, 08 2021.

Ali, M. and Khan, S. Clip-decoder : Zeroshot multilabel classification using multimodal clip aligned representation, 06 2024.

Guo, Y. Multimodal multilabel classification by clip, 06 2024.