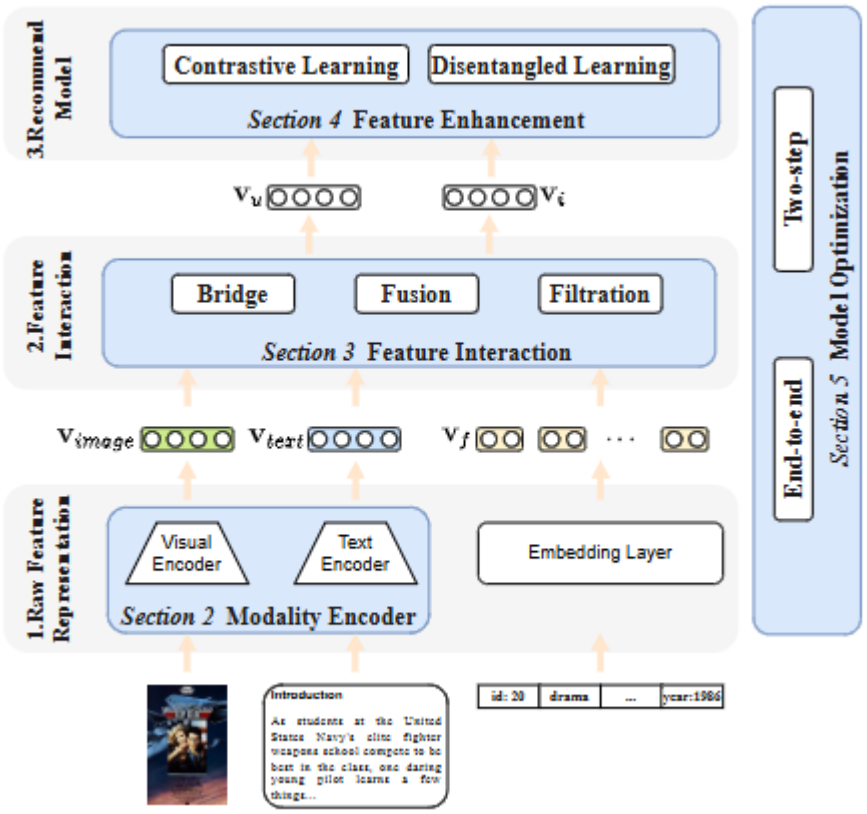
Multimodal Recommender Systems : A Survey

https://arxiv.org/pdf/2302.03883

0. Introduction

- 추천 시스템은 온라인 서비스에서 핵심 역할을 하며 사용자 선호도를 ID나 속성 기반으로 모델링함
- 멀티미디어 콘텐츠가 증가하면서 이미지와 텍스트 등 멀티모달 정보를 이해하는 추천 시스템이 중요해짐
- 멀티모달 정보는 데이터 희소성 문제를 완화하는 데 도움이 됨
- 이 논문은 Multimodal Recommender Systems 연구를 기술적 관점에서 정리한 서베이임

1. Overview



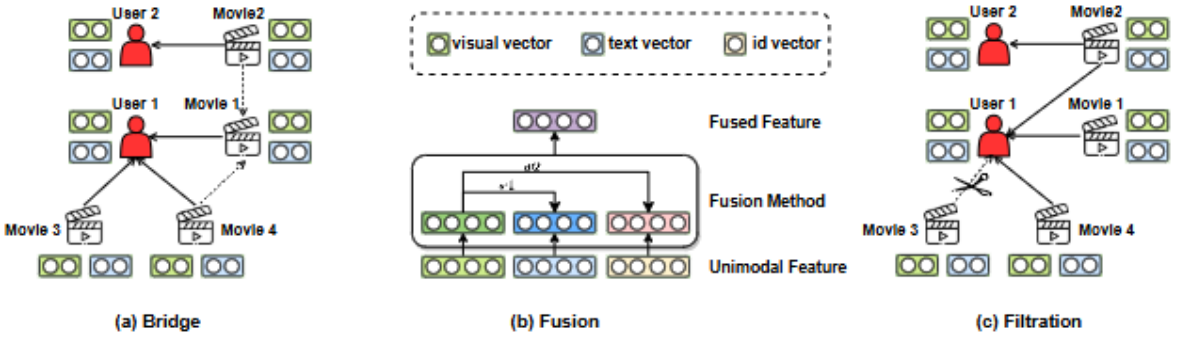
- MRS의 일반 절차는 세 단계로 구성됨
  - Raw Feature Representation
  - Feature Interaction
  - Recommendation Model
- 기술 분류
  - Modality Encoder
  - Feature Interaction
  - Feature Enhancement
  - Model Optimization
- 다양한 데이터셋과 코드 리소스를 함께 정리함
- MRS 연구의 흐름을 전체적으로 조망할 수 있게 구성됨

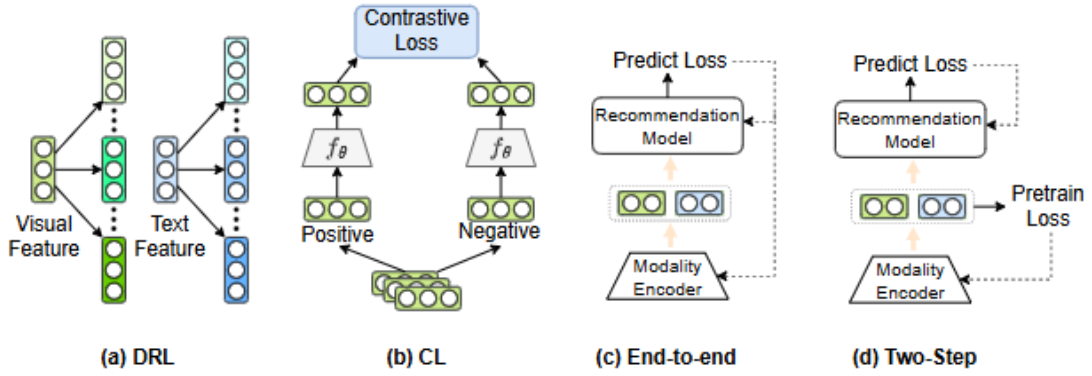
2. Challenges

- 데이터 희소성

- 모달별 표현 학습의 어려움
- 의미적 상호작용 모델링의 복잡성
- 멀티모달 피처 최적화 난이도
- 높은 계산 비용과 연산 자원 부담
- 데이터셋 및 구현체의 불일치로 인한 비교 연구 어려움

3. Method





- Modality Encoder
  - 이미지 : CNN
  - 텍스트 : Transformer, BERT, RNN 등
  - 기타 모달 : 도메인 특화 인코더
- Feature Interaction
  - 단순 융합

    - Attention 기반 Cross-modal Interaction
    - GNN 기반 관계 모델링
- Feature Enhancement
    - Self-supervised 학습
    - 데이터 증강
    - 보조 손실 및 정규화
- Model Optimization
    - 손실 함수 설계
    - 샘플링 전략
    - 정규화와 안정성 향상 기법
- 리소스 정리
    - 영화, 뉴스, 쇼핑 등 다양한 도메인의 멀티모달 추천 데이터셋 제공
    - 공개 코드 저장소 목록 제공

4. Experiments

Data	Field	Modality	Scale	Link
Tiktok	Micro-video	V,T,M,A	726K+	https://paperswithcode.com/dataset/tiktok-dataset
Kwai	Micro-video	V,T,M	1 million+	https://zenodo.org/record/4023390#.Y9YZ6XZBw7c
Movielens + IMDB	Movie	V,T	100k~25m	https://grouplens.org/datasets/movielens/
Douban	Movie,Book,Music	V,T	1 million+	https://github.com/FengZhu-Joey/GA-DTCDR/tree/main/Data
Yelp	POI	V,T,POI	1 million+	https://www.yelp.com/dataset
Amazon	E-commerce	V,T	100 million+	https://cseweb.ucsd.edu/ jmcauley/datasets.html#amazon_reviews
Book-Crossings	Book	V,T	1 million+	http://www2.informatik.uni-freiburg.de/ cziegler/BX/
Amazon Books	Book	V,T	3 million	https://jmcauley.ucsd.edu/data/amazon/
Amazon Fashion	Fashion	V,T	1 million	https://jmcauley.ucsd.edu/data/amazon/
POG	Fashion	V,T	1 million+	https://drive.google.com/drive/folders/1xFdx5xuNXHGsUVG2VIohFTXf9S7G5veq
Tianmao	Fashion	V,T	8 million+	https://tianchi.aliyun.com/dataset/43
Taobao	Fashion	V,T	1 million+	https://tianchi.aliyun.com/dataset/52
Tianchi News	News	T	3 million+	https://tianchi.aliyun.com/competition/entrance/531842/introduction
MIND	News	V,T	15 million+	https://msnews.github.io/
Last.FM	Music	V,T,A	186 k+	https://www.heywhale.com/mw/dataset/5cfe0526e727f8002c36b9d9/content
MSD	Music	T,A	48 million+	http://millionsongdataset.com/challenge/

[1] 'V', 'T', 'M', 'A' indicate the visual data, textual data, video data and acoustic data, respectively.

- 자체 실험이 아니라 기존 연구들의 데이터셋과 실험 구조 정리
- 주요 데이터 유형

    - 이미지 + 텍스트 조합
    - 제품 설명 + 시각 정보
    - 뉴스 기사 텍스트 + 이미지
- 비교 연구의 평가 지표와 실험 세팅을 체계적으로 정리함

5. Results

- 정량 실험은 없지만 기술 비교 분석 중심
- 기술별 장단점 정리
    - 단순 융합은 구현이 간단하지만 표현력이 부족함
    - Attention 기반 융합은 성능 우수하지만 비용이 큼
- MRS 연구의 주요 경향성 파악
- 공개 데이터·코드를 통해 연구 접근성을 크게 향상시킴

6. Insight

- MRS 연구를 체계적으로 조망할 수 있는 유용한 서베이
- 복잡한 기술을 네 가지 카테고리로 명확히 구조화
- 연구자가 모델을 설계하거나 비교할 때 참고할 수 있는 로드맵 역할
- 최신 LLM 기반 멀티모달 추천 논의는 부족함
- 실제 산업 환경에서의 비용·효율성 논의 부족
- 후속 연구 가능성
    - LLM 기반 MRS
    - 효율적 Cross-modal fusion
    - 대규모 실서비스 최적화
    - Explainability와 Privacy 관련 연구