모달리티 반영 뷰를 활용하는 대조 학습 기반의 멀티미디어 추천 시스템

반소희 [1], 김태리 [2], 김상욱 [3*]
[1] 한양대학교 미래모빌리티학과 석사과정
[2] 한양대학교 컴퓨터소프트웨어학과 박사과정
[3] 한양대학교 컴퓨터소프트웨어학과 교수

soheeb@hanyang.ac.kr, taerik@hanyang.ac.kr, wook@hanyang.ac.kr

Multimedia Recommender System Based on Contrastive Learning with Modality-Reflective View

SoHee Ban[1], Taeri Kim[2], Sang-Wook Kim[2*]
[1]Dept. of Future Mobility, Hanyang University
[2]Dept. of Computer Science, Hanyang University

요 약

최근, 대조 학습 기반의 멀티미디어 추천 시스템들이 활발하게 연구되고 있다. 이들은 아이템의 다양한 모달리티 피처들을 활용하여 사용자와 아이템에 대한 임베딩들(뷰들)을 생성하고, 이들을 통해 대조 학습을 진행한다. 학습한 뷰들을 추천에 활용함으로써, 이들은 기존 멀티미디어 추천 시스템들보다 상당히 향상된 추천 정확도를 획득했다. 그럼에도 불구하고, 우리는 기존 대조 학습 기반의 멀티미디어 추천 시스템들이 아이템의 뷰들을 생성하는 데에 아이템의 모달리티 피처들을 올바르게 반영하는 것의 중요성을 간과하며, 그 결과 추천 정확도 향상에 제약을 갖는다고 주장한다. 이는 아이템 임베딩에 아이템 자신의 모달리티 피처를 올바르게 반영하는 것이 추천 정확도에 향상에 도움이 된다는 기존 멀티미디어 추천 시스템의 발견에 기반한다. 따라서 본 논문에서 우리는 아이템의 모달리티 피처들을 올바르게 반영할 수 있는 뷰(구체적으로, 모달리티 반영 뷰)를 통해 대조 학습을 진행하는 새로운 멀티미디어 추천 시스템을 제안한다. 제안 방안은 두 가지 실세계 공개 데이터 집합들에 대해 최신 멀티미디어 추천 시스템보다 6.78%까지 향상된 추천 정확도를 보였다.

1. 서론

정보 과부화 시대에 추천 시스템은 사용자가 선호하는 아이템들을 쉽게 찾을 수 있도록 돕는 매우 중요한 역할을 하고 있다. 협업 필터링(Collaborative Filtering, 이하 CF)[1-7]은 이러한 추천 시스템들 중 대표적인 하나의 방법이지만, 매우 희소한 사용자와 아이템 간의 상호작용 정보(예: 클릭 로그, 구매 이력)만을 기반으로 동작하기 때문에(즉, 데이터 희소성 문제로 인해) 사용자의 선호를 정확하게 포착하기 어렵다는 한계를 가진다.

이러한 한계를 완화하기 위해, 사용자와 아이템 간의 상호작용 정보와 함께 아이템의 다양한 모달리티 피처들(예: 이미지 모달리티 피처, 텍스트 모달리티 피처)들을 추가로 활용하는 멀티미디어 추천 시스템[8-13]이 등장했다. 멀티미디어 추천 시스템들은 사용자와 아이템 간의 상호작용 정보 만으로는 알아내기 어려운, 각 모달리티에서 사용자가 선호하는 피처를 포착한 뒤 이러한 피처를 사용자가 선호할 법한 아이템들을 예측하는 데에 추가로 활용함으로써 기존 CF 방법들의 추천 정확도를 상당히 향상시켰다.

각 모달리티에서 사용자가 선호하는 피처를 포착하기 위해, 대부분의 멀티미디어 추천 시스템들은 먼저 아이템의 다양한 모달리티 피처들을 활용하여 사용자와 아이템에 대한 초기 임베딩들을 생성했다. 그러고 나서, 이들은 행렬 분해(Matrix Factorization), 그래프 합성곱 신경망(Graph Convolutional Network, 이하 GCN) [14]과 같은 방법을 활용해서 사용자와 아이템에 대한 임베딩들을 풍부하게 만든 뒤, 사용자와 아이템에 대한 임베딩들을 사용자와 아이템 간의 상호작용 정보를 기반으로 학습했다.

최근, 데이터 희소성 문제를 더욱 완화하고자 대조

* 교신 저자

학습(Contrastive Learning [15])을 추가로 활용하는 멀티미디어 추천 시스템들 [16-21]이 등장하고 있다. 대조 학습 기반의 멀티미디어 추천 시스템들은 아이템의 다양한 모달리티 피처들을 활용하여, 모달리티마다 사용자에 대한 뷰와 아이템에 대한 뷰를 생성한다. 그러고 나서, 이들은 같은 사용자(resp. 아이템)에 대한 뷰들끼리는 유사하게, 서로 다른 사용자들(resp. 아이템들)에 대한 뷰들끼리는 멀어지게 학습한 뒤, 학습한 뷰들을 추천에 활용한다. 뷰들을 통해 다양한 관점에서 풍부하게 학습된 사용자와 아이템의 정보를 활용함으로써, 대조 학습은 멀티미디어 추천 시스템들이 데이터 희소성 문제를 더욱 완화할 수 있도록 도움이 되었고, 그 결과 대조 학습 기반의 멀티미디어 추천 시스템들은 상당히 향상된 추천 정확도를 획득할 수 있었다. 따라서, 아이템의 다양한 모달리티 피처들을 활용하여 사용자와 아이템의 뷰들을 효과적으로 생성하는 것은 대조 학습 기반의 멀티미디어 추천 시스템들의 주된 관심사가 되어 왔다.

한편, (대조 학습을 활용하지 않는) 한 멀티미디어 추천 시스템 연구 [12]는 사용자와 아이템 간의 상호작용 정보 만을 기반으로 사용자와 아이템에 대한 임베딩들을 학습함으로써, 아이템이 자신의 고유한 피처를 담고 있는 모달리티 피처를 자신의 임베딩에 올바르게 반영하지 못하고 있다는 것 발견했다. 추가로, 해당 연구는 아이템의 임베딩에 아이템 자신의 모달리티 피처를 올바르게 반영하는 것이 추천 정확도 측면에서도 중요하다는 사실을 발견했다.

이러한 발견을 기반으로, 우리는 기존 대조 학습 기반의 멀티미디어 추천 시스템들이 아이템의 뷰를 생성하는 데에 아이템의 모달리티 피처를 반영하려고 시도조차 하지 않았다는 점을 지적한다. 따라서, 본 논문에서 우리는 아이템의 뷰를 생성하는 데에, 아이템의 모달리티 피처를 올바르게 반영할 수 있는 모달리티 반영 뷰를 통해 대조 학습을 진행하여 사용자에게 가장 정확한 추천을 제공하는 새로운 멀티미디어 추천 시스템을 제안하고자 한다.

2. 관련 연구

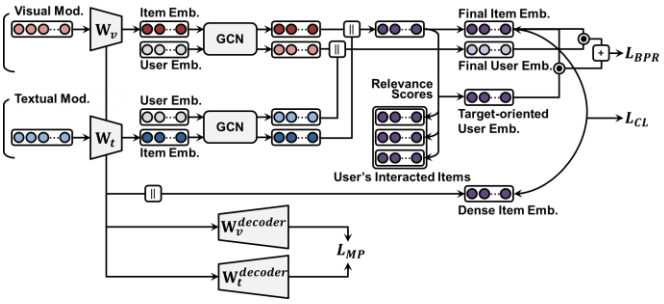
앞선 장에서 이야기한 바와 같이, 대조 학습 기반의 멀티미디어 추천 시스템들의 주된 관심사는 효과적인 뷰의 생성이다. 따라서 우리는 본 장에서 기존 방법들이 뷰를 어떻게 생성하는지에 대해 구체적으로 이야기하고자 한다.

MICRO [18]는 아이템의 다양한 모달리티 피처들을 기반으로 모달리티마다 아이템-아이템 그래프를 구축한 뒤, 구축된 그래프들에 GCNs 를 적용함으로써 아이템들 간의 관계를 반영하는 아이템 뷰들을 생성했다. MMGCL [17]은 모달리티마다 사용자-아이템 상호작용 이분 그래프를 구축한 뒤, 구축된 그래프마다 랜덤으로 에지들을 드롭하는 모달리티 에지 드롭아웃 방법과, 구축된 그래프들 중 랜덤으로 선택된 한 모달리티에 대한 그래프를 마스킹(다시 말하자면, 모든 에지들을 드롭아웃)하는 모달리티 마스킹 방법을 적용함으로써 사용자의 모달리티 선호를 구별하는 사용자와 아이템의 뷰들을 생성했다.

그러나, 아이템의 임베딩에 아이템 자신의 모달리티 피처를 올바르게 반영하는 것이 추천 정확도 측면에서도 중요함에도 불구하고, 이들은 아이템의 뷰를 생성하는 데에 노드들 간의 관계를 반영하는 것에만 관심을 두었고, 아이템의 모달리티 피처를 반영하려는 시도조차 하지 않았다.

3. 제안 방안

본 장에서 우리는 아이템의 모달리티 피처들을 올바르게 반영할 수 있는 뷰(즉, 모달리티 반영 뷰)를 통한 대조 학습 기반의 새로운 멀티미디어 추천 시스템을 제안한다. 이것의 오버뷰는 (그림 1)과 같다.



(그림 1) 제안 방안의 오버뷰

먼저 제안 방안은 모달리티마다 사용자-아이템 상호작용 이분 그래프를 구축한다. 이때, 사용자 임베딩은 모달리티마다 다르게 랜덤한 값으로 초기화하고, 아이템 임베딩은 다층 퍼셉트론을 통해 아이템의 모달리티 피처로부터 획득한 덴스한 임베딩으로 초기화한다. 이후, 모달리티마다 독립적인 GCN [13]을 적용하여 모달리티 별 사용자와 아이템 임베딩들을 학습한 뒤, 모달리티 별 사용자와 아이템 임베딩들을 사용자와 아이템마다 순차(concatenation)하여 최종 사용자 임베딩과 최종 아이템 임베딩을 획득한다.

이제, 우리는 아이템의 모달리티 피처를 올바르게 반영할 수 있는 모달리티 반영 뷰를 통해 모달리티마다 아이템의 뷰를 획득하고, 대조 학습을 진행한다. 멀티미디어 추천 시스템들의 학습 과정에서 아이템의 모달리티 피처가 손실되는 부분은 다층 퍼셉트론을 바탕으로 아이템의 모달리티 피처를 압축한 뒤 이를 아이템 임베딩으로 설정하는 부분(즉, 초기화 부분)이기 때문에, 우리는 복원을 위한 다층 퍼셉트론을 추

가로 구축하고 오토인코더 기반의 Modality Preservation (이하 MP) 로스 함수 [12]를 활용한다. MP 로스는 모달리티마다 압축된 아이템 임베딩과 복원된 아이템 임베딩(즉, 뷰) 간의 차이를 최소화함으로써 아이템의 뷰에 아이템의 모달리티 피처를 올바르게 반영하게 한다.

추가로, 우리는 사용자가 한 아이템(이하 타겟 아이템)에 대한 상호작용 여부를 결정할 때, 사용자가 과거에 상호작용했던 아이템들 중, 모달리티 피처 측면에서 타겟 아이템과 유사한 아이템들에 대한 선호가 타겟 아이템에 대한 상호작용 여부에 더 많은 영향을 미칠 것이라는 직관을 고려하기 위해, 어텐션 네트워크 [13]를 기반으로 타겟 아이템에 대한 구체적인 선호를 나타내는 추가적인 사용자 임베딩(이하 타겟 지향 사용자 임베딩)을 획득한다.

최종적으로, 최종 사용자 임베딩과 타겟 지향 사용자 임베딩, 그리고 아이템 자신의 모달리티 피처를 올바르게 반영한 최종 아이템 임베딩 간의 dot product를 기반으로 타겟 아이템에 대한 사용자의 예측 선호도를 계산한 뒤, 각 사용자에 대해 예측 선호도가 높은 아이템들을 추천한다.

4. 실험

4-1. 실험 환경

<표 1> 데이터 셋 통계치

데이터 집합	사용자수	아이템수	상호작용수	희소성
Baby	19,445	7,050	160,792	99.88%
Toys and Games	19,412	11,924	167,597	99.93%

본 장에서 우리는 제안 방안의 효과를 검증하기 위해, 멀티미디어 추천 시스템 연구들 [11-13, 16, 18, 21]에서 널리 사용되는 실세계 공개 데이터 집합들인 Amazon Baby, Amazon Toys and Games [22]의 두 가지 카테고리 데이터 집합들과, 최신 멀티미디어 추천 시스템들 [11-13, 17, 18, 21]을 사용하여 실험을 진행한다.

<표 1>은 각 데이터 집합에 대한 구체적인 통계치를 나타낸다. 각 데이터 집합에는 사용자와 아이템 간의 상호작용 정보와 각 아이템의 이미지 모달리티 피처(4,096 차원), 텍스트 모달리티 피처(1,024 차원)가 모두 포함되며, 각 사용자와 각 아이템은 최소 5 개의 상호작용을 가진다. 우리는 기존 멀티미디어 추천 시스템들 [9-13, 16-21]과 동일하게, 각 데이터 집합에 대해 각 사용자의 상호작용 중 80%를 랜덤으로 선택하여 학습 집합으로, 나머지 20% 중 절반은 검증 집합으로, 나머지 절반은 실험 집합으로 구성했다.

우리는 제안 방안과 경쟁 방안들의 추천 정확도를 계산하기 위해, 모든 방안들에 대해서 사용자마다 예측 선호도가 가장 높은 10 개의 아이템들을 추천한 뒤, 멀티미디어 추천 시스템 연구들 [9-13, 16-21]에서 널리 사용되고 있는 평가 지표인 Precision, Recall, NDCG 를 사용했다.

4-2. 실험 결과

<표 2> Baby 데이터 집합에서
제안 방안과 경쟁 방안들의 추천 정확도

	Precision@10	Recall@10	NDCG@10
FREEDOM	0.0058	0.0552	0.0291
LATTICE	0.0053	0.0506	0.0276
MARIO	0.0048	0.0467	0.0245
MMGCL	0.0053	0.0505	0.0267
MONET	0.0059	0.0564	0.0310
MICRO	0.0057	0.0547	0.0304
제안 방안	**0.0063**	**0.0600**	**0.0329**
개선 (%)	**6.78**	**6.38**	**6.13**

<표 3> Toys and Games 데이터 집합에서
제안 방안과 경쟁 방안들의 추천 정확도

	Precision@10	Recall@10	NDCG@10
FREEDOM	0.0095	0.0899	0.0490
LATTICE	0.0096	0.0919	0.0534
MARIO	0.0088	0.0849	0.0488
MMGCL	0.0087	0.0813	0.0449
MONET	0.0112	0.1059	0.0623
MICRO	0.0098	0.0932	0.0532
제안 방안	**0.0116**	**0.1088**	**0.0638**
개선 (%)	**3.57**	**2.74**	**2.41**

실험 결과는 <표 2>, <표 3>에 나타나 있으며, 데이터 집합과 평가 지표마다 가장 높은 추천 정확도는 굵게, 두 번째로 높은 추천 정확도는 밑줄로 표시되어 있다. 추가로, 가장 높은 추천 정확도를 보이는 방안이 두 번째로 높은 추천 정확도를 보이는 방안의 추천 정확도를 얼마나 개선하였는지를 백분율로 계산하여 표에 함께 나타냈다.

제안 방안은 모든 데이터 집합들에 대해 모든 경쟁 방안들보다 일관적으로 높은 추천 정확도를 보였다. 구체적으로, 제안 방안은 가장 높은 추천 정확도를 보이는 경쟁 방안인 MONET [13]를 Baby, Toys and Games 데이터 집합에 대해 추천 정확도를 각각 최대 6.78%, 3.57%까지 향상시켰다. 추가로, 제안 방안은 아이템 임베딩에 자신의 모달리티 피처를 올바르게 반영하고자 한 MARIO [12]를 Baby, Toys and Games 데이터 집합에 대해 추천 정확도를 각각 최대 34.29%, 31.82%까지 향상시켰다.

이는 대조 학습의 이점을 최대한 활용하면서도, 모달리티 반영 뷰를 활용하여 아이템 뷰에 아이템 자신의 모달리티 피처를 올바르게 반영하는 것이 추천 정확도 향상에 상당히 효과적임을 뒷받침한다.

5. 결론

본 논문에서 우리는 아이템의 뷰를 생성하는 데에, 아이템의 모달리티 피처를 올바르게 반영할 수 있는 모달리티 반영 뷰를 통해 대조 학습을 진행하여 사용자에게 가장 정확한 추천을 제공하는 새로운 멀티미디어 추천 시스템을 제안했다. 두 가지 실세계 공개 데이터 집합들과 최신 멀티미디어 추천 시스템들을 활용한 실험들을 통해 우리는 주장의 유효성과 제안 방안의 효과를 검증했다.

사사

이 논문은 2024 년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원과 한국연구재단의 지원을 받아 수행된 연구임(No.2022-0-00352, No.RS-2022-00155586, No.2018R1A5A7059549)

참고문헌

[1] Chae, Dong-Kyu, et al. "Rating augmentation with generative adversarial networks towards accurate collaborative filtering." The World Wide Web Conference. 2019.

[2] Chae, Dong-Kyu, et al. "AR-CF: Augmenting virtual users and items in collaborative filtering for addressing cold-start problems." Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval. 2020.

[3] Kong, Taeyong, et al. "Linear, or non-linear, that is the question!." Proceedings of the fifteenth ACM international conference on web search and data mining. 2022.

[4] Lim, Hongjun, et al. "AiRS: a large-scale recommender system at naver news." 2022 IEEE 38th International Conference on Data Engineering (ICDE). IEEE, 2022.

[5] Rendle, Steffen, et al. "BPR: Bayesian personalized ranking from implicit feedback." arXiv preprint arXiv:1205.2618 (2012).

[6] Su, Yixin, et al. "Neural graph matching based collaborative filtering." Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval. 2021.

[7] Chen, Lei, et al. "Set2setRank: Collaborative set to set ranking for implicit feedback based recommendation." Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2021.

[8] He, Ruining, and Julian McAuley. "VBPR: visual bayesian personalized ranking from implicit feedback." Proceedings of the AAAI conference on artificial intelligence. Vol. 30. No. 1. 2016.

[9] Wei, Yinwei, et al. "MMGCN: Multi-modal graph convolution network for personalized recommendation of micro-video." Proceedings of the 27th ACM international conference on multimedia. 2019.

[10] Wei, Yinwei, et al. "Graph-refined convolutional network for multimedia recommendation with implicit feedback." Proceedings of the 28th ACM international conference on multimedia. 2020.

[11] Zhang, Jinghao, et al. "Mining latent structures for multimedia recommendation." Proceedings of the 29th ACM international conference on multimedia. 2021.

[12] Kim, Taeri, et al. "MARIO: modality-aware attention and modality-preserving decoders for multimedia recommendation." Proceedings of the 31st ACM International Conference on Information & Knowledge Management. 2022.

[13] Kim, Yungi, et al. "MONET: Modality-Embracing Graph Convolutional Network and Target-Aware Attention for Multimedia Recommendation." Proceedings of the 17th ACM International Conference on Web Search and Data Mining. 2024.

[14] Kipf, Thomas N., and Max Welling. "Semi-supervised classification with graph convolutional networks." arXiv preprint arXiv:1609.02907 (2016).

[15] Chen, Ting, et al. "A simple framework for contrastive learning of visual representations." International conference on machine learning. PMLR, 2020.

[16] Yu, Penghang, et al. "Multi-view graph convolutional network for multimedia recommendation." Proceedings of the 31st ACM International Conference on Multimedia. 2023.

[17] Yi, Zixuan, et al. "Multi-modal graph contrastive learning for micro-video recommendation." Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2022.

[18] Zhang, Jinghao, et al. "Latent structure mining with contrastive modality fusion for multimedia recommendation." IEEE Transactions on Knowledge and Data Engineering (2022).

[19] Tao, Zhulin, et al. "Self-supervised learning for multimedia recommendation." IEEE Transactions on Multimedia (2022).

[20] Wei, Yinwei, et al. "Contrastive learning for cold-start recommendation." Proceedings of the 29th ACM International Conference on Multimedia. 2021.

[21] Zhou, Xin, and Zhiqi Shen. "A tale of two graphs: Freezing and denoising graph structures for multimodal recommendation." Proceedings of the 31st ACM International Conference on Multimedia. 2023.

[22] McAuley, Julian, et al. "Image-based recommendations on styles and substitutes." Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval. 2015.