양방향 Mamba와 어텐션 풀링을 결합한 순차 추천 모델

박민선, 김정현
세종대학교
22012146@sju.ac.kr, j.kim@sejong.ac.kr

A Sequential Recommendation Model Combining Bidirectional Mamba and Attention Pooling

Minseon Park, Junghyun Kim
Sejong Univ.

요 약

본 논문에서는 순차 추천 시스템에서 발생하는 정보 손실과 문맥 포착의 한계를 극복하기 위해 양방향 맘바와 어텐션 풀링을 결합한 새로운 모델 구조를 제안한다. 제안 모델은 사용자 행동 시퀀스를 순방향과 역방향의 이중 경로로 처리하여 양방향 의존성을 학습하고, 어텐션 메커니즘을 통해 각 시점의 중요 정보를 선별적으로 통합한다. 실험 결과, 제안 모델은 기존 모델 대비 파라미터 수와 연산량이 다소 증가하였으나, 순위 정확도가 최대 13.76% 향상되는 성과를 보였다.

Ⅰ. 서 론

사용자의 과거 행동 이력을 바탕으로 다음 행동을 예측하는 순차 추천 시스템은 전자상거래 및 콘텐츠 플랫폼의 핵심 기술이다[1]. 기존 Transformer 기반 모델의 연산 효율성 문제를 해결하기 위해 선형 복잡도를 가진 선택적 상태 공간 모델(SSMs)인 Mamba[2]가 등장하였고, 이를 순차 추천에 적용한 Mamba4Rec[3]은 효율적인 추론 속도를 입증하였다. 그러나 Mamba4Rec은 단방향 구조로 인해 시퀀스 전반의 전역적 문맥(Global Context)을 충분히 반영하는 데 한계가 있으며, 마지막 시점의 정보만을 활용하는 Last-State Pooling 방식으로 인해 초기 정보가 소실되는 정보 병목(Information Bottleneck) 문제가 존재한다.

본 논문에서는 이러한 한계를 해결하기 위해 양방향 맘바 구조와 어텐션 풀링을 결합한 새로운 모델을 제안한다. 제안 모델은 SIGMA[4]의 구조화 전략에 착안하여, 관측된 사용자 행동 시퀀스를 순방향과 역방향으로 재구성하여 각각 인코딩하였다. 이를 통해 현재 시점까지의 과거 행동들 간의 상대적 문맥 관계를 보다 풍부하게 반영한다. 또한, SASRec[5] 등에서 검증된 어텐션 메커니즘을 풀링 단계에 적용하여 시퀀스 내 각 시점의 중요도를 선별적으로 반영함으로써 정보 손실을 최소화하였다. 공개 데이터셋을 활용한 실험 결과, 제안 모델은 기존 Mamba4Rec 대비 일관된 성능 향상을 보이며 표현력을 향상시킨 순차 추천 모델임을 입증하였다.

Ⅱ. 본론

본 논문에서는 순차 추천에서 단방향 모델이 가지는 문맥 표현의 한계를 극복하기 위해 양방향 Mamba 구조와 어텐션 풀링을 결합한 새로운 추천 모델을 제안한다. 제안 모델은 사용자의 과거 행동 시퀀스를 입력으로 받아 다음 시점에서의 아이템 선호도를 예측하는 것을 목표로 하며, 전체 구조는 그림 1과 같이 임베딩 층, 다층의 양방향 Mamba 인코더, 그리고 어텐션 풀링 층으로 구성된다.

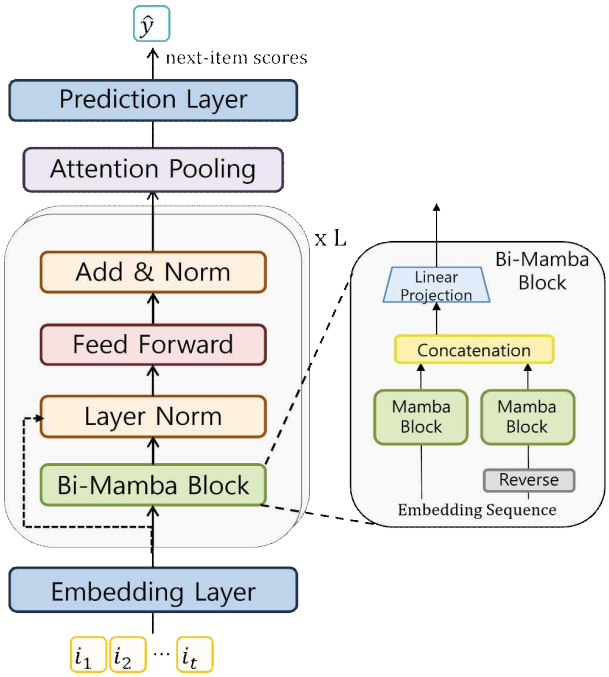

그림 1. 제안하는 양방향 맘바 및 어텐션 풀링 결합 모델 구조

먼저, 사용자 u의 아이템 상호작용 시퀀스 $S_u = [i_1, i_2, i_3, ..., i_t]$는 임베딩 층을 통해 고정 차원의 연속적인 벡터 시퀀스로 변환된다. 이 과정에서 각 아이템은 잠재 공간에서의 의미를 갖는 임베딩 벡터로 표현되며, 이후 Dropout과 Layer Normalization을 적용하여 모델 학습의 안정성과 일반화 성능을 향상시킨다. 이렇게 생성된 임베딩 시퀀스는 제안하는 Mamba 인코더의 입력으로 사용된다.

제안 모델의 핵심 구성 요소인 양방향 Mamba 인코더는 기존 단방향

Mamba 기반 모델의 한계를 보완하기 위해 이중 경로 구조로 설계되었다. 구체적으로, 입력 시퀀스는 순방향과 역방향의 두 경로로 처리된다. 순방향 경로에서는 사용자 행동의 시간적 흐름에 따라 $[i_1 \rightarrow i_t]$방향으로 Mamba 블록을 적용하여 인과적 문맥을 학습하며, 역방향 경로에서는 동일한 과거 시퀀스를 역순으로 재배열하여 $[i_t \rightarrow i_1]$방향으로 Mamba 블록을 적용한다. 이는 미래 정보를 참조하는 것이 아니라 동일한 과거 데이터에 대해 회고적 문맥을 학습하는 방식으로 추천 시점에서의 정보 누수를 방지한다.

각 경로에서 출력된 은닉 표현은 결합 연산을 통해 하나의 시퀀스 표현으로 통합되며, 선형 투영을 통해 차원이 조정된다. 이후 잔차 연결과 Layer Normalization을 적용하여 원본 임베딩 정보가 효과적으로 보존되도록 하였으며, Feed-Forward Network를 통해 비선형 변환을 수행함으로써 표현력을 강화하였다. 이러한 양방향 Mamba 레이어를 다층으로 구성함으로써 모델은 시퀀스 전반에 걸친 전역적 문맥 정보를 효과적으로 학습할 수 있다.

양방향 Mamba 인코더를 통과한 시퀀스 표현은 어텐션 풀링 층으로 전달된다. 기존의 Last-item Pooling 방식은 마지막 시점의 정보에만 의존하기 때문에 추천에 중요한 과거 행동 정보가 충분히 반영되지 않는 한계를 가진다. 이를 해결하기 위해 본 논문에서는 학습 가능한 쿼리 기반의 어텐션 풀링을 도입하였다. 어텐션 풀링 층에서는 시퀀스 내 각 시점의 은닉 상태에 대해 중요도를 계산하고, Softmax 함수를 통해 정규화된 가중치를 산출한다. 이후 이 가중치를 적용한 가중합을 통해 시퀀스 전체를 대표하는 사용자 벡터를 생성하며 Layer Normalization을 적용하여 표현의 안정성을 확보한다. 이를 통해 모델은 다음 아이템 예측에 기여하는 핵심 행동 시점을 선택적으로 강조할 수 있다.

최종적으로 생성된 사용자 표현 벡터는 전체 아이템 임베딩과의 내적 연산을 통해 다음 아이템에 대한 선호 점수를 계산한다. 모델 학습에는 Cross-Entropy Loss 또는 BPR Loss를 사용하였으며, 이를 통해 실제 사용자 선택과 모델 예측 간의 차이를 최소화하도록 최적화하였다.

제안 모델의 성능을 검증하기 위해 순차 추천의 대표적인 벤치마크인 MovieLens-1M과 Amazon-Video-Games 데이터셋을 사용하여 실험을 수행하였다. 베이스라인 모델로는 단방향 구조인 Mamba4Rec[3]을 선정하여 양방향 처리와 어텐션 풀링의 효과를 비교 분석하였다. 모든 실험 설정은 공정한 비교를 위해 선행 연구와 동일하게 구성하였다.

본 논문에서는 순차 추천 성능을 평가하기 위해 상위 10개 아이템에 대한 Hit@10, NDCG@10, MRR@10을 사용하였다. 먼저, Hit@10은 추천된 상위 10개의 아이템 목록 내에 실제 사용자가 상호작용한 정답 아이템이 포함되었는지를 측정하여 모델의 아이템 탐색 성능을 평가한다. 반면, NDCG@10과 MRR@10은 정답 아이템의 순위 위치를 고려하여 추천 결과의 순위 정확도를 평가한다. NDCG@10은 정답 아이템이 상위에 위치할수록 더 높은 점수를 부여하며, MRR@10은 정답 아이템이 처음 등장하는 순위의 역수를 평균하여 최상위 추천 결과의 정확도를 반영한다. 이러한 지표들을 종합적으로 활용함으로써 추천 성공 여부와 순위 품질을 함께 평가하고자 한다.

실험 결과, 제안 모델은 두 데이터셋 모두에서 기존 모델 [3] 대비 일관된 성능 향상을 보였다. 특히 ML-1M 데이터셋에서 MRR@10과 NDCG@10이 각각 13.76%, 11.75% 향상되었으며, Amazon 데이터셋에서는 Hit@10이 8.26% 증가하였다. 비록 이중 경로 학습으로 인해 기존 모델 대비 파라미터 수가 증가하였지만, 추천 정확도와 순위 품질 측면에서 일관된 성능 향상을 보였다. 이에 대한 자세한 결과는 표 1과 표 2에 제시되어 있다.

표 1. 기존 모델[3]과 제안 모델의 성능 비교

Datasets	Eval Metrics	기존 모델[3]	**제안 모델**	Improv.
MovieLens-1M	Hit@10	0.3109	**0.3374**	8.80%
	NDCG@10	0.1795	**0.2006**	11.75%
	MRR@10	0.1395	**0.1587**	13.76%
Amazon-Video-Games	Hit@10	0.1150	**0.1245**	8.26%
	NDCG@10	0.0612	**0.0629**	2.78%
	MRR@10	0.0450	0.0444	-1.33%

표 2. 기존 모델[3]과 제안 모델의 복잡도 비교

Method	GPU memory	Parameters
기존 모델 [3]	6.02G	290944
제안 모델	6.86G	457344

Ⅲ. 결론

본 논문에서는 순차 추천 시스템의 정보 병목 현상을 해결하고 전역적 문맥 포착 능력을 강화하기 위해 양방향 맘바와 어텐션 풀링을 결합한 새로운 모델을 제안하였다. 제안 모델은 사용자 행동 시퀀스를 입력받아 순방향 및 역방향의 이중 경로로 재구성한 뒤, 이를 양방향 맘바 인코더에서 병렬로 처리하여 시퀀스 내의 인과적 흐름과 회고적 문맥을 동시에 학습하였다. 이후 두 경로의 출력을 결합하고 어텐션 풀링을 적용하여 각 시점의 중요 정보를 선별적으로 통합함으로써 최종 사용자 선호도를 정교하게 예측하였다. 실험 결과, 제안 모델은 기존 모델 대비 파라미터 수와 메모리 용량이 다소 증가하였으나 추천 정확도와 순위 품질 측면에서 일관된 성능 향상을 보였다. 향후에는 연산 효율성을 개선하기 위해 인코더 간의 가중치 공유 설계를 도입하거나 모델 경량화 기법을 적용하고, 이를 통해 성능 저하 없이 구조를 최적화하여 실제 대규모 서비스 환경에서도 효율적으로 동작할 수 있도록 개선하고자 한다.

ACKNOWLEDGMENT

이 논문은 2023년도 정부 (교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임 (RS-2023-00271991).

참 고 문 헌

[1] S. Wang, L. Hu, Y. Wang, L. Cao, Q. Z. Sheng, and M. Orgun, "Sequential Recommender Systems: Challenges, Progress and Prospects," in Proc. Int. Joint Conf. Artif. Intell. (IJCAI), 2019, pp. 6332–6338.

[2] A. Gu and T. Dao, "Mamba: Linear-time sequence modeling with selective state spaces," arXiv preprint arXiv:2312.00752, 2023.

[3] C. Liu, J. Lin, J. Wang, H. Liu, and J. Caverlee, "Mamba4rec: Towards efficient sequential recommendation with selective state space models," arXiv preprint arXiv:2403.03900, 2024.

[4] Z. Liu, Q. Liu, et al., "SIGMA: Selective Gated Mamba for Sequential Recommendation," arXiv preprint arXiv:2408.11451, 2024.

[5] W. C. Kang and J. McAuley, "Self-Attentive Sequential Recommendation," in Proc. IEEE Int. Conf. Data Mining (ICDM), 2018, pp. 197–206.