Towards More Accurate Sequential Recommendations

313551075 Chen Shih-Hong[1] 313551179 Yu Jun-You[1] 110550085 Fang Tien-Yueh[1]

[1]National Yang Ming Chiao Tung University

Abstract

Sequential recommendation systems seek to predict a user's next interaction by modeling the evolving order of past behaviors. [1] defines the "training data development" problem and aims to obtain optimal training data by learning a new dataset that explicitly captures item transition patterns. Building on this and inspired by [2], we further design a "retrieval-augmented module" to fully leverage the regenerated data from [1] for training the sequential recommendation model. Our training pipeline can be divided into three stages: (i) Dataset Regeneration, (ii) Recommendation Learning, and (iii) Retrieval-augmented Fine-tuning. We evaluate the effectiveness of our approach on three publicly available datasets suitable for sequential recommendation: Amazon Beauty, Toys, and Sports.

Method

Our work is divided into three main steps:

1. **Dataset Regeneration:** Following [1], we train a dataset regenerator, and then perform hybrid inference to generate a new training dataset that captures item transition patterns.
2. **Recommendation Learning:** Any sequential recommendation model can serve as the backbone; here, we use [3] and [4] to learn user behavior from the regenerated dataset.
3. **Retrieval-augmented Fine-tuning:** We construct a memory bank using two different methods: (1) Enumerating all collaborative `<user sequence, target item>` pairs from training data, and (2) Mining frequent sequential patterns as references. These pairs are encoded into vector embeddings and stored. When a user input sequence is given, the model retrieves the top-K similar user representations from the memory bank. Using the current user vector as the query, the retrieved user vectors as keys, and the corresponding item embeddings as values, a Multi-Head Cross Attention (MHCA) module generates an augmented representation:

$$h' = \mathrm{MHCA}\big(h,\ \{h_k\}_{k=1}^{K},\ \{v_k\}_{k=1}^{K}\big)$$

The final user representation is computed as a weighted combination:

$$\tilde{h} = \alpha\, h + (1-\alpha)\, h',$$

where h is the original user embedded vector, h′ is the augmented vector from MHCA, and $\alpha \in [0,1]$ controls the balance. The model is then fine-tuned with a softmax loss:

$$\mathcal{L}_{\mathrm{raft}} = -\log\Big(\frac{e^{\tilde{h}^T v_{t+1}}}{\sum_{v_i \in \mathcal{V}} e^{\tilde{h}^T v_i}}\Big)$$

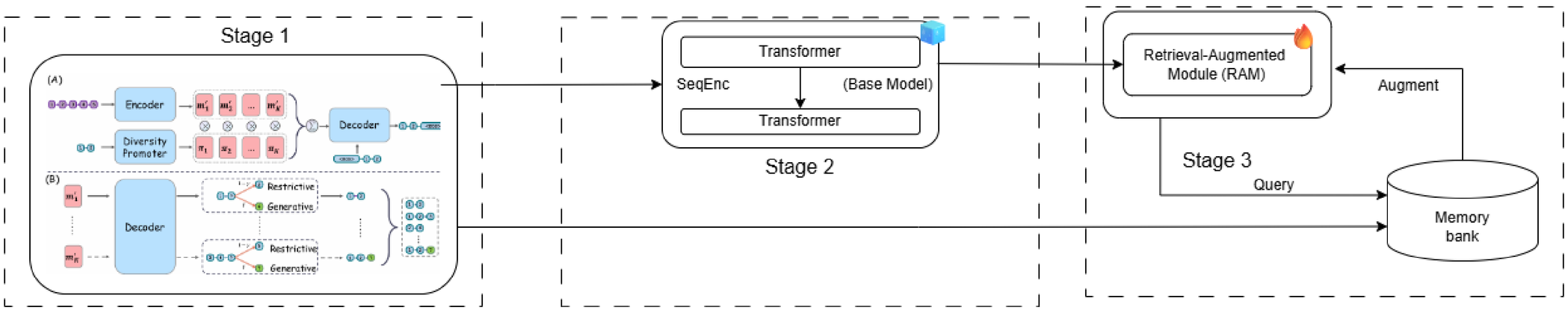
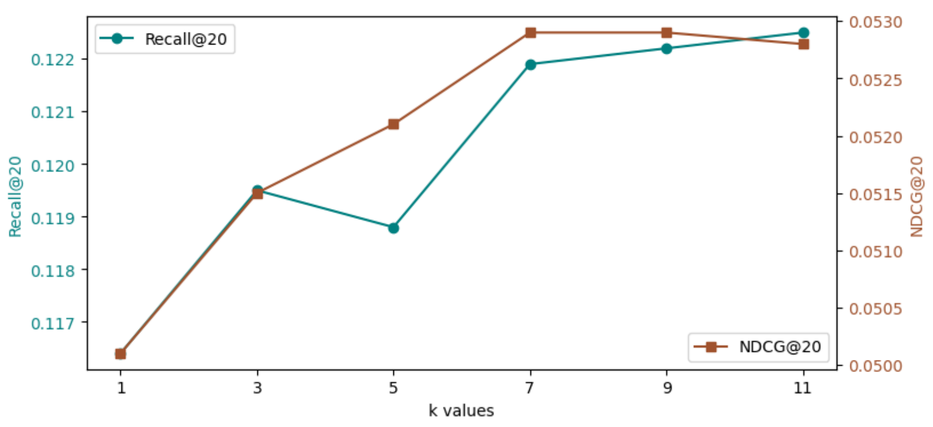


Figure 1. Training pipeline

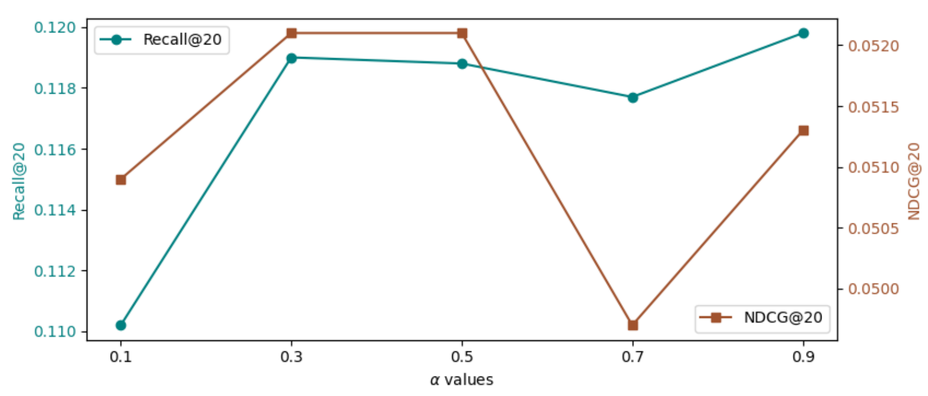
Experiment Result

Table 1. Performance comparison on Amazon Beauty, Toys, and Sports datasets. R@k denotes Recall@k and N@k denotes NDCG@k. "Improvement" shows the relative percent gain of "(Ours)" over the base method.

	Beauty						Toys						Sports					
Method	R@5	R@10	R@20	N@5	N@10	N@20	R@5	R@10	R@20	N@5	N@10	N@20	R@5	R@10	R@20	N@5	N@10	N@20
SASRec	0.0555	0.0819	0.1162	0.0319	0.0405	0.0491	0.0619	0.0890	0.1225	0.0349	0.0436	0.0520	0.0311	0.0472	0.0691	0.0170	0.0222	0.0277
SASRec (Ours)	**0.0578**	**0.0844**	**0.1188**	**0.0349**	**0.0434**	**0.0521**	**0.0641**	**0.0925**	**0.1250**	**0.0370**	**0.0461**	**0.0543**	0.0307	0.0482	**0.0719**	0.0174	0.0231	**0.0290**
Improvement	4.14%	3.05%	2.24%	9.40%	7.16%	6.11%	3.55%	3.93%	2.04%	6.02%	5.73%	4.42%	-1.29%	2.12%	4.05%	2.35%	4.05%	4.69%
DuoRec	0.0570	0.0860	0.1212	0.0359	0.0452	0.0541	0.0634	0.0934	0.1269	0.0380	0.0476	0.0561	0.0301	0.0453	0.0667	0.0192	0.0241	0.0294
DuoRec (Ours)	**0.0578**	**0.0864**	**0.1229**	**0.0378**	**0.0470**	**0.0562**	**0.0654**	**0.0946**	**0.1292**	**0.0403**	**0.0497**	**0.0584**	**0.0310**	**0.0464**	**0.0686**	**0.0203**	**0.0253**	**0.0308**
Improvement	1.40%	0.47%	1.40%	5.29%	3.98%	3.88%	3.15%	1.28%	1.81%	6.05%	4.41%	4.10%	2.99%	2.43%	2.85%	5.73%	4.98%	4.76%



(a) Beauty (varying k)



(b) Beauty (varying α)

Figure 2. Recommendation performance w.r.t. different k and α on Amazon Beauty dataset.

Discussion and Conclusion

Experiments on Amazon Beauty, Toys, and Sports datasets show that our retrieval-augmented module consistently improves the performance of sequential recommendation models (SASRec, DuoRec). However, since our method already incorporates contrastive learning, the gain may diminish when the base model (e.g. DuoRec) also uses similar techniques. Effectively integrating retrieval augmentation with contrastive-based models remains a potential direction for future work.

References

[1] M. Yin, H. Wang, W. Guo, Y. Liu, S. Zhang, S. Zhao, D. Lian, and E. Chen, "Dataset Regeneration for Sequential Recommendation," *arXiv:2405.17795*, 2024.

[2] X. Zhao, B. Hu, Y. Zhong, S. Huang, Z. Zheng, M. Wang, H. Wang, and M. Zhang, "RaSeRec: Retrieval-Augmented Sequential Recommendation," *arXiv:2412.18378*, 2024.

[3] W.-C. Kang and J. McAuley, "Self-Attentive Sequential Recommendation," *arXiv:1808.09781*, 2018.

[4] R. Qiu, Z. Huang, H. Yin, and Z. Wang, "Contrastive Learning for Representation Degeneration Problem in Sequential Recommendation," *arXiv:2110.05730*, 2021.