RESETBERT4Rec: A Pre-training Model Integrating Time And User Historical Behavior for Sequential Recommendation

Qihang Zhao
zhaoqihang3@jd.com
University of Science and Technology of China, Hefei, Anhui, China
JD AI Research, Shanghai, China

ABSTRACT

Sequential recommendation methods are very important in modern recommender systems because they can well capture users' dynamic interests from their interaction history, and make accurate recommendations for users, thereby helping enterprises succeed in business. However, despite the great success of existing sequential recommendation-based methods, they focus too much on item-level modeling of users' click history and lack information about the user's entire click history (such as click order, click time, etc.). To tackle this problem, inspired by recent advances in pre-training techniques in the field of natural language processing, we build a new pre-training task based on the original BERT pre-training framework and incorporate temporal information. Specifically, we propose a new model called the **RE**arrange **S**equence pr**E**-training and **T**ime embedding model via **BERT** for sequential **R**ecommendation (**RESETBERT4Rec**) [1], it further captures the information of the user's whole click history by adding a rearrange sequence prediction task to the original BERT pre-training framework, while it integrates different views of time information. Comprehensive experiments on two public datasets as well as one e-commerce dataset demonstrate that RESETBERT4Rec achieves state-of-the-art performance over existing baselines.

CCS CONCEPTS

• **Information systems** → **Social recommendation**;

KEYWORDS

Sequential Recommendation, Rearrange Sequence Prediction, Pre-training

ACM Reference Format:
Qihang Zhao. 2022. RESETBERT4Rec: A Pre-training Model Integrating Time And User Historical Behavior for Sequential Recommendation. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '22), July 11–15, 2022, Madrid, Spain.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3477495.3532054

[1]This work was completed during JD internship.

SIGIR '22, July 11–15, 2022, Madrid, Spain

ACM ISBN 978-1-4503-8732-3/22/07…$15.00
https://doi.org/10.1145/3477495.3532054

1 INTRODUCTION

Nowadays, recommender systems have been widely employed in major online platforms, such as Amazon, Facebook, Twitter, etc., to meet the interests and requirements of users. However, with the development of Internet technology, information has begun to explode rapidly, and on these online platforms, the interests and requirements of users are constantly changing and evolving dynamically and rapidly over time. However, with the development of Internet technology, information begins to explode rapidly. On these online platforms, users' interests and requirements change and evolve dynamically and rapidly over time, which makes it more and more difficult for the online platform to make appropriate recommendations for users. To cope with this problem and mine the interests hidden behind users' click behaviors, scholars have proposed a variety of methods for sequential recommendation [1, 2, 11, 14, 16, 22].

Typically, methods for sequential recommendation all aim to mine the sequential behavior patterns of users by mining the user's click behavior history. With the development of deep learning, such motivation has been widely attempted based on deep learning models. Various methods for sequential recommendation adopt different deep learning models: recurrent neural networks model (RNNs) [7], convolutional neural networks model (CNNs) [15], Transformer [14] and so on to learn the proper representation of user according to the interaction of user with item.

Futhermore, some researchers have also realized that rich contextual information (such as item attributes) can significantly improve the performance of sequential recommender systems [8, 10, 23]. Simultaneously, with the development of computer vision and natural language processing, some scholars began to apply some technologies in the field of CV and NLP, such as pre-training technology and self supervised learning, to the sequential recommendation system, and achieved good results. Specifically, BERT4Rec [14] utilizes the BERT model to build a bidirectional sequence model to learn the representation of user from the user's historical interaction behavior, and then generate the appropriate next recommendation item for the user in the sequential recommendation task. S^3_*Rec* [24] constructs self-supervised signals to improve the representation of users in sequential recommendation by paying attention to the intrinsic correlation of data. CP4Rec [20] adopts the method of mask and crop sequence items to construct positive and negative samples, uses the contrastive learning paradigm to extract meaningful user behavior patterns, and alleviates the problem of sample data sparsity.

Despite these sequential recommendation methods have shown excellent performance in sequential recommendation tasks, we argue that these sequential recommendation methods still have

Figure 1: Example of click history behavior sequence of two users on an online platform.

some shortcomings. Firstly, the existing sequential recommendation methods often pay too much attention to a single item in the user interaction history, but cannot fully extract the information of the whole sequence level of the user interaction history, such as the user's click order. As shown in Figure 1, subfigure (a) and subfigure (b) are the browsing history sequences of two users of an online platform for the category of electric toothbrushes. The items they browse are almost the same, but the click order is different, which shows that the click order of users' items can reflect users' preferences to a certain extent. Actually, there has been some researches [9, 18] on user click order, but it still cannot make good use of order information. Another major limitation is that the existing sequence recommendation methods ignore the interaction timestamp value between users and items. Timestamp values are quite useful in sequential click tasks because they rife with information. Therefore, if the timestamp information can be introduced into the sequential recommendation model, it will have a great gain for the sequential recommendation task.

To address the defects mentioned above, Specifically, inspired by the pre-training mechanism proposed by BERT [4] and enabling BERT to succeed in text understanding, we first propose a new pre-training task for sequential recommendation task based on BERT: rearrange sequence prediction, to learn the whole sequence information of user interaction history behavior. Specifically, we rearrange the user's original interaction history sequence according to a certain probability, and let the model predict whether the input sequence is rearranged. In this way, the model can learn the corresponding order preference through the whole interaction behavior of users. At the same time, the rearrange sequence prediction can better learn the coarse-grained (sequence-level) and fine-grained (item-level) information of user interaction history together with the close task focusing on a single item. Furthermore, to better utilize timestamp values, we propose a multi-view temporal encoding mechanism to mine user behavior patterns from different perspectives. Extensive experiments show that our proposed RESETBERT4Rec significantly outperforms other baseline methods on two public datasets and one e-commerce dataset.

The contributions of our paper are as follows:

- We propose a novel pre-training task for sequential recommendation: rearrange sequence prediction, which learns users' preferences from the whole sequence level of users. In addition, we also design a multi view time coding scheme to model the time information of user interaction behavior.
- We conducted comprehensive experiments on two public sequential recommendation data sets and one e-commerce data set to show the effectiveness of our proposed RESETBERT4Rec.
- We also perform comprehensive ablation study on RESETBERT4Rec to help understand the role of each key component in the sequential recommendation task.

2 PRELIMINARIES

In this section, we first define the sequential recommendation problem, and then introduce the language model BERT.

2.1 Problem Definition

In order recommendation, let $\mathcal{U} = \{u_1, u_2, \ldots, u_{|U|}\}$ represent the set of U users and $\mathcal{I} = \{i_1, i_2, \ldots, i_{|M|}\}$ represent the set of M items, and list $\mathcal{H}_u = [i_1^{(u)}, i_2^{(u)}, \ldots, i_t^{(u)}, \ldots, i_{n_u}^{(u)}]$ represents the interaction history sequence of user u in chronological order. For user $u \in \mathcal{U}$, $i_t^{(u)} \in \mathcal{I}$ is the item that user u interacted with at time stamp t, and n_u is the last timestamp of user interaction history. Then the sequential recommendation task can be defined as: for a user $u \in \mathcal{U}$, given him/her interaction history sequence $\mathcal{H}_u$, predict the item that the user u will interact with at $n_u + 1$. It can be formalized as the probability of modeling all possible items of user u in time step $n_u + 1$:

$$\mathcal{P}(i_{n_u+1}^{(u)} | \mathcal{H}_u) \tag{1}$$

2.2 BERT

Since our RESETBERT4Rec is based on BERT, we need to briefly introduce BERT before we introduce our model. BERT is the state-of-the-art pre-training language model. The main module is the stack of the encoder part of the Transformer [17] model. Its design purpose is to jointly conditioning the left and right contexts in all layers, through the two pre-training tasks of mask LM and Next Sentence Prediction, pre-training deep bidirectional representations from unlabeled text. Thus, pre-trained BERT models can be fine-tuned with just one additional output layer to create state-of-the-art models for a wide range of tasks, such as question answering and language inference.

Given a text sequence $\mathcal{S}$, the main pipeline of Bert is to embed the text sequence: token embedding, segment embedding and position embedding, and then add these three types of embedding into the stacked transformer encoder to obtain the output, and then use the output to optimize the model by training two pre-training tasks. It can be formalized as:

$$o = TrE(TrE(\ldots TrE(TE(\mathcal{S}) + SE(\mathcal{S}) + PE(\mathcal{S})))) \tag{2}$$

Where o represents the output, TrE represents the encoder of the Transformer, TE represents the token embedding, SE represents the segment embedding, and PE represents the position embedding. Due to the powerful performance of BERT in text sequence processing, some scholars naturally migrate it to sequential recommendation task [14, 20, 21], and achieved excellent results.

3 MODEL

In this section, we will introduce RESETBERT4Rec in detail, and our model is developed based on BERT. The procedures of its pre-training and fine-tuning stages are shown in the Figure 2.

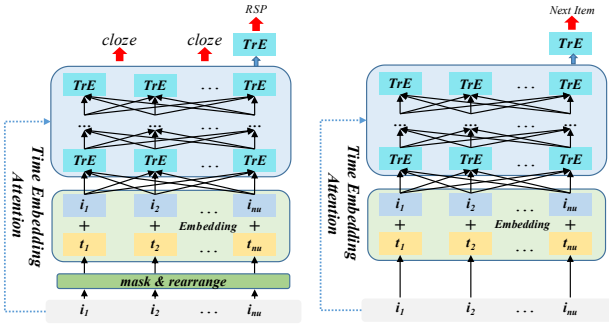

Figure 2: Overall pre-training and fine-tuning procedures for RESETBERT4Rec.

3.1 Embedding Layer

For the embedding layer, compared to BERT, we replace the fixed sine-cosine position embedding with a learnable relative time interval embedding. Specifically, for a given item i_t at timestamp t, its input representation h_i^0 is constructed by summing the corresponding item and relative time interval embedding:

$$h_i^0 = IE(i_t) + TE(RTI(t)) \tag{3}$$

Where IE is the item embedding table, TE is the time interval embedding table, and RTI is the relative time interval conversion of timestamps: $RTI(t) = \lfloor n * (t - t_1)/(t_{nu} - t_1) \rfloor$, where t_1 and t_{nu} represent the timestamps of the start and end of the sequence, respectively, and n (integer) is the specified number of time intervals, here we take $n = 20$.

3.2 Pre-training RESETBERT4Rec

Pre-training Task: The pre-training stage is shown in Figure 2 (a).

Task #1: Cloze Prediction: Following the previous research on sequential recommendation [14], we also follow the cloze task in our pre-training task, that is, the item id in a given sequence is randomly masked with probability p_1, and after model processing, the final hidden vector corresponding to the masked item is fed to the output softmax on the item set. Finally, we define the loss of each masked input $\mathcal{H}'_u$ as the negative log likelihood of the masked target:

$$\mathcal{L}_{cloze} = \frac{1}{|\mathcal{H}|} \sum_{j \in \mathcal{H}_u} -logP(i_t^{(u)} = i_t^{(u*)} | \mathcal{H}'_u) \tag{4}$$

where $\mathcal{H}$ is the browsing history of all users, $\mathcal{H}'_u$ is the masked interaction sequence, and $i_t^{(u*)}$ is the true value of the masked item $i_t^{(u)}$.

Task #2: Rearrange Sequence Prediction (RSP): As mentioned above, we propose a new auxiliary task in this work: Rearrange Sequence Prediction. We note that, as shown in Figure 1, for a user, the item interaction order in his/her interaction history is an important factor in capturing the user's true preferences. It is worth mentioning that in previous work, there are also some studies [13, 19, 20] that reorder the user's interaction sequence. However, most of these methods use reordering as a data augmentation technique to construct positive samples of original samples under the framework of comparative learning, without noticing that the order information of user interaction history sequence is used to capture user preference. very critical. Inspired by this, this task aims to rearrange the items of the user interaction history with a certain probability, and use the model to predict whether the user interaction history has been rearranged to learn the sequence-level information of the user's entire interaction history.

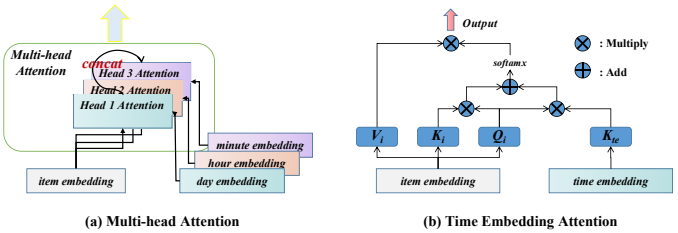

Figure 3: Details of time embedding attention.

Specifically, given a sequence of user interactions $\mathcal{H}_u = [i_1^{(u)}, i_2^{(u)}, i_3^{(u)}, i_4^{(u)}, i_5^{(u)}]$, we reorder the items of the sequence with probability p_2, the rearranged sequence is $\widetilde{\mathcal{H}_u} = [i_2^{(u)}, i_4^{(u)}, i_1^{(u)}, i_5^{(u)}, i_3^{(u)}]$ (Note that this is just one case of all possible rearrangement scenarios). In rearrange sequence prediction task, for an interaction sequence that has not been rearranged, its label is '`unrearranged`', otherwise it is '`rearranged`'. In other words, we treat rearrange sequence prediction as a binary classification task, so we can define the loss for the rearrange sequence prediction task as:

$$\mathcal{L}_{RSP} = -\frac{1}{|\mathcal{H}|} \sum_{u \in \mathcal{U}} \sum_{j=0,1} y_{uj} p(\widetilde{y_{uj}} | \widetilde{\mathcal{H}_u}) \tag{5}$$

Pre-training: In the pre-training stage, for the cloze task, we directly take the output of BERT as the final representation vector. For the RSP task, in order to further extract the sequence-level information, we superimpose a transformer encoder at the end of BERT and sum its output as the final representation vector. RESERBERT4Rec trains the model by jointly optimizing Loss 4 and Loss 5 in order to adapt it to downstream tasks for the subsequent fine-tuning stage.

Multi-view Time Embedding Attention: In order to better capture the user's interest preferences, we introduce a multi-view time embedding attention mechanism in both the pre-training and fine-tuning stages, as shown in Figure 3. Details as follow:

To further model temporal information, we provide embeddings from three different perspectives based on timestamps of user interaction history: day embeddings, hourly embeddings, and minute embeddings, and feed different embeddings to different heads of the transformer's multi-head attention mechanism , act as different experts to improve diversity for decision-making, as shown in Figure 3 (a). At the same time, in this paper, we do not use the practice of adding time embedding and item embedding and then perform self attention, but adopt the practice of previous research [3], as shown in Figure 3 (b).

3.3 Fine-tuning RESETBERT4Rec

We have pre-trained RESERBERT4REC in the previous section, in order to adapt RESERBERT4REC to the sequential recommendation task, we need to fine-tune the model. Specifically, in the input stage, we no longer mask and rearrange the input sequence, and for output, we use the output module originally used for RSP task to predict the next item, and regard the entire sequential recommendation task as a multi-classification task.

4 EXPERIMENTS

4.1 Experimental Settings

Datasets. We use two public data sets: MovieLens [5] (**ML-1m**) and **Beauty**, and an e-commerce data set **JD-Etooth**: The interaction history of users on the category of electric toothbrushes collected from JD platform spans

Table 1: Statistics of datasets.

Datasets	ML-1m	Beauty	JD-Etooth
#Users	6040	40226	311018
#Items	3416	54542	15799
#Actions	1.0m	0.35m	4.8m
#Avg.length	163.5	8.8	15.6

from July 16, 2021 to July 20, 2021. The statistics of the data set are listed in Table 1.

Evaluation. The last two items in each item sequence are used for validation and testing purposes, respectively, and the remaining items are used for model training. Performance is evaluated by Recall@5, NDCG@5, Recall@10 and NDCG@10. To reduce the time consumption of evaluation, following the strategy in [14], we randomly sample 99 negative items for each true item. Then, each metric is calculated based on the ranking of these items, and the average metric score on the test data is reported.

Baselines. To validate the effectiveness of our RESETBERT4Rec, we compare it with state-of-the-art baselines. Specifically, first we choose the classic **NCF** [6], **GRU4Rec** [7] and **SASRec** [11] as the comparison methods. Second, since our RESETBERT4REC is developed based on BERT, we choose a BERT-based methods: **BERT4Rec** [14]. Finally, since **TiSASRec** [12] also successfully integrates temporal information, we also choose it as one of the baseline methods.

Parameter Settings. First, for hyper-parameters common across all models, we consider hidden dimension d of $\{16, 32, 64, 128, 256\}$, the l_2 regularizer considers $\{1, 0.1, 0.01, 0.001, 0.001, 0.0001\}$, and dropout rate from $\{0, 0.1, 0.2, , 0.9\}$. Secondly, for the unique hyper-parameters in the baseline methods such as **GRU4Rec** [7], **SASRec** [11], **BERT4Rec** [14], and **TiSASRec** [12], we directly follow the optimal parameter configuration given in the corresponding paper, and we report the results of each baseline under its optimal hyper-parameter setting.

Second, for RESETBERT4REC, all parameters are initialized with a truncated normal distribution in the range $[-0.02, 0.02]$, we train the model with Adam with a learning rate of $1e-4$, β_1=0.9, β_2=0.999, and the l_2 weight decay is 0.0, the learning rate decays linearly. When the l_2 norm of a gradient exceeds a threshold of 5, the gradient is clipped. We set the layer number L=2 and the head number h=2, and for the head setting, we empirically set the dimension of each head to 32. The mask probability p_1 of JD-Etooth and ML-1m is 0.15, while the mask probability p_1 of Beauty is 0.4. The rearrangement probability p_2 is 0.2. We set the maximum input sequence to 50 for the Beauty and JD-Etooth datasets and 200 for ML-1m. All models are trained from scratch without any pre-training on a single NVIDIA Tesla V100 GPU with a batch size of 128.

4.2 Experimental Results

The performance of RESETBERT4Rec and baseline models are shown in Table 2. From the comparison results, we can see that RESETBERT4Rec outperforms the corresponding base models on all three datasets in terms of all evaluation metrics. Such results show that the RSP task and the time embedding attention mechanism we designed are able to capture the user's behavior pattern well, and then can achieve excellent performance in the sequential recommendation task. Specifically, for TiSASRec, although it successfully integrates the timestamp information of user interaction history, its performance is worse than RESETBERT4Rec, which shows the effectiveness of our proposed time embedding attention mechanism. Meanwhile, BERT4Rec shows the best performance among all the baseline methods, but its performance is inferior to RESETBERT4Rec, which also shows the importance of user interaction history sequence level information.

We performed an ablation study to observe the effectiveness of the proposed RESETBERT4Rec. Due to space limitations, Table 3 only reports the

Table 2: Performance comparison of different methods on next-item prediction. 'RESET' is short for RESETBERT4Rec.

Dataset	Baseline	Recall@5	Recall@10	NDCG@5	NDCG@10
ML-1m	NCF	0.1897	0.3479	0.1142	0.1637
	GRU4Rec	0.4674	0.6205	0.3182	0.3631
	SASRec	0.5435	0.6621	0.3982	0.4365
	TiSASRec	0.5869	0.7103	0.4302	0.4701
	BERT4Rec	0.5877	0.6987	0.4432	0.4801
	RESET	**0.6312**	**0.7428**	**0.4728**	**0.5213**
Beauty	NCF	0.1308	0.2140	0.0853	0.1125
	GRU4Rec	0.1315	0.2344	0.0813	0.1075
	SASRec	0.1935	0.2654	0.1437	0.1632
	TiSASRec	0.2082	0.2806	0.1513	0.1746
	BERT4Rec	0.2208	0.3026	0.1608	0.1872
	RESET	**0.2503**	**0.3345**	**0.1927**	**0.2103**
JD-Etooth	NCF	0.1602	0.2813	0.0944	0.1273
	GRU4Rec	0.1709	0.3166	0.0967	0.1289
	SASRec	0.2421	0.3250	0.1903	0.2188
	TiSASRec	0.2663	0.3425	0.2178	0.2365
	BERT4Rec	0.2991	0.4532	0.1821	0.2191
	RESET	**0.3329**	**0.5021**	**0.2213**	**0.2407**

Table 3: Ablation Study of RESETBERT4Rec on JD-Etooth Dataset.

Baseline	Recall@5	Recall@10	NDCG@5	NDCG@10
w/o time	0.3074	0.4629	0.1914	0.2231
w/o RSP	0.3257	0.4873	0.2147	0.2362
RESETBERT4Rec	**0.3329**	**0.5021**	**0.2213**	**0.2407**

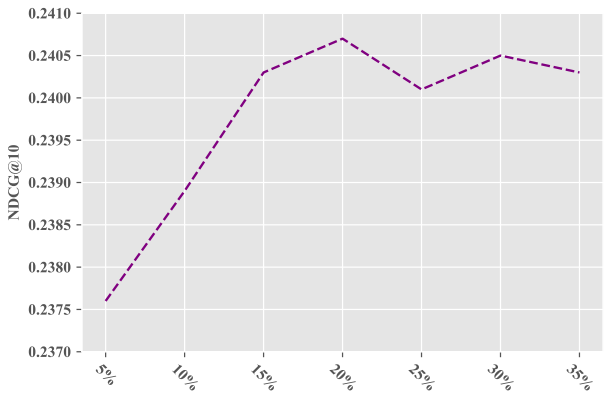

Figure 4: Sensitivity of RESETBERT4Rec over p_2 on JD-Etooth Dataset.

experimental results of RESETBERT4Rec on JD-Etooth. It can be seen that both RSP and time embedding attention help to improve RESETBERT4Rec. And the role of time embedding attention is slightly larger than that of RSP.

Finally, we investigate the sensitivity of RESETBERT4Rec to the rearrangement probability p_2. Figure 4 shows the NDCG@10 scores of RESETBERT4Rec with different p_2. The performance reaches the best performance when p_2 is around 20%, which is the reason we set p_2 to 20% in the above

experiments. Simultaneously, with the increase of p_2, the performance of model does not change much, that is, the model is no longer sensitive to p_2.

5 CONCLUSION

In this paper, we propose a novel pre-training model for sequential recommendation: RESETBERT4Rec, which can learn the sequence-level information of user interaction history through a rearrange sequence prediction task designed by us, and we also propose a novel time embedding attention mechanism, which can well model the timestamp information of user interaction history. Under the combined effect of the two, the user's behavior patterns and interest preferences can be better captured according to the user's interaction history. Extensive experiments show that our RESETBERT4Rec outperforms other baselines on the sequential recommendation task.

ACKNOWLEDGEMENT

This work was supported by the National Key R&D Program of China [Grant number 2020AAA0103804] and the National Natural Science Foundation of China [Grant number 72004021]..

REFERENCES

[1] Guohao Cai, Xiaoguang Li, Quanyu Dai, Gang Wang, Zhenhua Dong, Chaoliang Zhang, Xiuqiang He, and Lifeng Shang. 2021. Dual Sequence Transformer for Query-based Interactive Recommendation. In *22nd IEEE International Conference on Mobile Data Management, MDM 2021, Toronto, ON, Canada, June 15-18, 2021*. IEEE, 139–144. https://doi.org/10.1109/MDM52706.2021.00030

[2] Renqin Cai, Jibang Wu, Aidan San, Chong Wang, and Hongning Wang. 2021. Category-aware Collaborative Sequential Recommendation. In *SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, July 11-15, 2021*. ACM, 388–397. https://doi.org/10.1145/3404835.3462832

[3] Zihang Dai, Zhilin Yang, Yiming Yang, Jaime G. Carbonell, Quoc Viet Le, and Ruslan Salakhutdinov. 2019. Transformer-XL: Attentive Language Models beyond a Fixed-Length Context. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*. Association for Computational Linguistics, 2978–2988. https://doi.org/10.18653/v1/p19-1285

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, 4171–4186. https://doi.org/10.18653/v1/n19-1423

[5] F. Maxwell Harper and Joseph A. Konstan. 2016. The MovieLens Datasets: History and Context. *ACM Trans. Interact. Intell. Syst.* 5, 4 (2016), 19:1–19:19. https://doi.org/10.1145/2827872

[6] Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. 2017. Neural Collaborative Filtering. In *Proceedings of the 26th International Conference on World Wide Web, WWW 2017, Perth, Australia, April 3-7, 2017*. ACM, 173–182. https://doi.org/10.1145/3038912.3052569

[7] Balázs Hidasi, Alexandros Karatzoglou, Linas Baltrunas, and Domonkos Tikk. 2016. Session-based Recommendations with Recurrent Neural Networks. In *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*. http://arxiv.org/abs/1511.06939

[8] Balázs Hidasi, Massimo Quadrana, Alexandros Karatzoglou, and Domonkos Tikk. 2016. Parallel Recurrent Neural Network Architectures for Feature-rich Session-based Recommendations. In *Proceedings of the 10th ACM Conference on Recommender Systems, Boston, MA, USA, September 15-19, 2016*. ACM, 241–248. https://doi.org/10.1145/2959100.2959167

[9] Liang Hu, Longbing Cao, Shoujin Wang, Guandong Xu, Jian Cao, and Zhiping Gu. 2017. Diversifying Personalized Recommendation with User-session Context. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017*. ijcai.org, 1858–1864. https://doi.org/10.24963/ijcai.2017/258

[10] Xiaowen Huang, Shengsheng Qian, Quan Fang, Jitao Sang, and Changsheng Xu. 2018. CSAN: Contextual Self-Attention Network for User Sequential Recommendation. In *2018 ACM Multimedia Conference on Multimedia Conference, MM 2018, Seoul, Republic of Korea, October 22-26, 2018*. ACM, 447–455. https://doi.org/10.1145/3240508.3240609

[11] Wang-Cheng Kang and Julian J. McAuley. 2018. Self-Attentive Sequential Recommendation. In *IEEE International Conference on Data Mining, ICDM 2018, Singapore, November 17-20, 2018*. IEEE Computer Society, 197–206. https://doi.org/10.1109/ICDM.2018.00035

[12] Jiacheng Li, Yujie Wang, and Julian J. McAuley. 2020. Time Interval Aware Self-Attention for Sequential Recommendation. In *WSDM '20: The Thirteenth ACM International Conference on Web Search and Data Mining, Houston, TX, USA, February 3-7, 2020*. ACM, 322–330. https://doi.org/10.1145/3336191.3371786

[13] Kyuyong Shin, Hanock Kwak, Kyung-Min Kim, Minkyu Kim, Young-Jin Park, Jisu Jeong, and Seungjae Jung. 2021. One4all User Representation for Recommender Systems in E-commerce. *CoRR* abs/2106.00573 (2021). arXiv:2106.00573 https://arxiv.org/abs/2106.00573

[14] Fei Sun, Jun Liu, Jian Wu, Changhua Pei, Xiao Lin, Wenwu Ou, and Peng Jiang. 2019. BERT4Rec: Sequential Recommendation with Bidirectional Encoder Representations from Transformer. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management, CIKM 2019, Beijing, China, November 3-7, 2019*. ACM, 1441–1450. https://doi.org/10.1145/3357384.3357895

[15] Qiaoyu Tan, Jianwei Zhang, Jiangchao Yao, Ninghao Liu, Jingren Zhou, Hongxia Yang, and Xia Hu. 2021. Sparse-Interest Network for Sequential Recommendation. In *WSDM '21, The Fourteenth ACM International Conference on Web Search and Data Mining, Virtual Event, Israel, March 8-12, 2021*. ACM, 598–606. https://doi.org/10.1145/3437963.3441811

[16] Jiaxi Tang and Ke Wang. 2018. Personalized Top-N Sequential Recommendation via Convolutional Sequence Embedding. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining, WSDM 2018, Marina Del Rey, CA, USA, February 5-9, 2018*. ACM, 565–573. https://doi.org/10.1145/3159652.3159656

[17] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*. 5998–6008. https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

[18] Shoujin Wang, Liang Hu, Longbing Cao, Xiaoshui Huang, Defu Lian, and Wei Liu. 2018. Attention-Based Transactional Context Embedding for Next-Item Recommendation. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*. AAAI Press, 2532–2539. https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16318

[19] Zhuofeng Wu, Sinong Wang, Jiatao Gu, Madian Khabsa, Fei Sun, and Hao Ma. 2020. CLEAR: Contrastive Learning for Sentence Representation. *CoRR* abs/2012.15466 (2020). arXiv:2012.15466 https://arxiv.org/abs/2012.15466

[20] Xu Xie, Fei Sun, Zhaoyang Liu, Jinyang Gao, Bolin Ding, and Bin Cui. 2020. Contrastive Pre-training for Sequential Recommendation. *CoRR* abs/2010.14395 (2020). arXiv:2010.14395 https://arxiv.org/abs/2010.14395

[21] Xu Yuan, Dongsheng Duan, Lingling Tong, Lei Shi, and Cheng Zhang. 2021. ICAI-SR: Item Categorical Attribute Integrated Sequential Recommendation. In *SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, July 11-15, 2021*. ACM, 1687–1691. https://doi.org/10.1145/3404835.3463060

[22] Shengyu Zhang, Dong Yao, Zhou Zhao, Tat-Seng Chua, and Fei Wu. 2021. CauseRec: Counterfactual User Sequence Synthesis for Sequential Recommendation. In *SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, July 11-15, 2021*. ACM, 367–377. https://doi.org/10.1145/3404835.3462908

[23] Tingting Zhang, Pengpeng Zhao, Yanchi Liu, Victor S. Sheng, Jiajie Xu, Deqing Wang, Guanfeng Liu, and Xiaofang Zhou. 2019. Feature-level Deeper Self-Attention Network for Sequential Recommendation. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*. ijcai.org, 4320–4326. https://doi.org/10.24963/ijcai.2019/600

[24] Kun Zhou, Hui Wang, Wayne Xin Zhao, Yutao Zhu, Sirui Wang, Fuzheng Zhang, Zhongyuan Wang, and Ji-Rong Wen. 2020. S3-Rec: Self-Supervised Learning for Sequential Recommendation with Mutual Information Maximization. In *CIKM '20: The 29th ACM International Conference on Information and Knowledge Management, Virtual Event, Ireland, October 19-23, 2020*. ACM, 1893–1902. https://doi.org/10.1145/3340531.3411954