RecFormer: personalized temporal-aware transformer for fair music recommendation

Wei-Yao Wang[1], Wei-Wei Du[1] and Wen-Chih Peng[1]

[1]*Department of Computer Science, National Yang Ming Chiao Tung University, Hsinchu, Taiwan*

Abstract

Recommendation systems have improved the characterization of user preferences by modeling their digital footprints and item content. However, another facet, model behavior, has attracted a great deal of attention in both academic and industry fields in recent years due to the increasing awareness of fairness. The shared task, a Rounded Evaluation of Recommender Systems (EvalRS @ CIKM 2022), is introduced to broadly measure multifaceted model predictions for music recommendation. To tackle the problem, we propose the RecFormer architecture with a personalized temporal-aware transformer to model the interactions among user history in a single framework. Specifically, RecFormer adopts a masked language modeling task as the training objective, which enables the model to capture fine-grained track embeddings by reconstructing tracks. Meanwhile, it also integrates a temporal-aware self-attention mechanism into the Transformer architecture so that the model is able to consider time-variant information among different users. Moreover, we introduce linearized attention to reduce quadratic computation and memory cost since the limited time is one of the challenges in this task. Extensive experiments and analysis are conducted to demonstrate the effectiveness of our RecFormer compared with the official baseline, and we examine the model contribution from the ablation study. Our team, yao0510, won the seventh prize with a total score of 0.1964 in the EvalRS challenge, which illustrates that our model achieved competitive performance. The source code will be publicly available at https://github.com/wywyWang/RecFormer.

Keywords

Recommendation System, User Fairness, Transformer, Temporal-Aware

1. Introduction

In recent years, characterizing users with accurate interests has evolved due to the use of advanced recommendation systems (RSs). These RSs have been used to develop several real-world applications in industry, for example, Amazon ROSE [1]. While there has been significant progress in predicting accurate items for users of RSs, the awareness of model behavior has attracted a great deal of attention from both academic and industry researchers. As recommendation systems are built on top of users, data, and models, it is likely that the system will make unfair suggestions due to the biases of these candidates [2], which illustrates the increasing need to investigate model behavior.

To that end, a rounded evaluation of recommender systems hosted by EvalRS @ CIKM 2022 is introduced to tackle both standard evaluation metrics and model behavior tests for music recommendation. Figure 1 illustrates an example of a music recommendation system. Given multiple past tracks from a set of target users, our recommendation system aims to predict a set of songs that are most likely to be listened to by the target users. On top of the standard retrieval metrics (i.e., hit-rate (HR), mean reciprocal rank (MRR), and normalized discounted cumulative gain (nDCG)), the organizers further employ multi-dimensional non-observational perspectives to evaluate model behaviors.

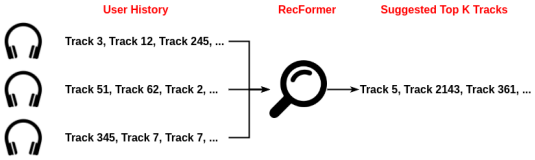

Figure 1: An example pipeline of the music recommendation with our proposed RecFormer.

However, we introduce that there are at least three challenges in addressing this shared task. 1) **Validation strategy**. The task uses bootstrapped nested cross-validation[1] that randomly samples a track from the user sequence as the test set, which means that this task cannot be formulated as a conventional sequence classification task that uses first N tracks to predict the $N+1$ track since the test set may not be selected from the latest track the user listened to. 2) **Time limitation**. The training and inference time is required to be less than

[1]https://github.com/RecList/evalRS-CIKM-2022/blob/main/images/loop.jpg

CIKM'22: Proceedings of the 31st ACM International Conference on Information and Knowledge Management

sf1638.cs05@nctu.edu.tw (W. Wang); wwdu.cs10@nycu.edu.tw (W. Du); wcpeng@cs.nycu.edu.tw (W. Peng)
https://wywywang.github.io/ (W. Wang); https://wwweiwei.github.io/ (W. Du)
0000-0002-6551-1720 (W. Wang); 0000-0002-0627-0314 (W. Du); 0000-0002-0172-7311 (W. Peng)

© 2022 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).
CEUR Workshop Proceedings (CEUR-WS.org)

CEUR-WS.org/Vol-3318/short9.pdf

22.5 minutes per fold, which is challenging when adopting deep learning techniques to have a fine-grained track representation if the space and computation complexity is high. 3) **Time-variant event**. The target users have their corresponding habits of listening music songs during the day; for example, students are likely to listen to music after school, while office workers are likely to listen during working hours. For example, Koren [3] is the first approach that verifies the effectiveness of modeling temporal effect in the Netflix competition. Therefore, converting user history into a sequence for recurrent-based approaches directly ignores the time-variant influence of listening to music tracks. Moreover, there is no existing evaluation metric that is able to measure the model behavior in terms of the time domain. It is essential to take temporal information into account and design a proper metric for the corresponding evaluation.

To address the aforementioned challenges, we propose **RecFormer**, a novel personalized temporal-aware Transformer for fair music recommendation, which consists of personalized user embeddings, a temporal-aware multi-head linearized-attention in a modified Transformer [4], and a track classifier to predict the possible tracks. Specifically, the personalized user embeddings take rich user-related metadata into account to represent each user. To tackle the first challenge, we employ the masked language modeling task as our training objective, which randomly masks a proportion of all tracks in the input sequence. For the second and third challenges, we introduced a temporal-aware linearized-attention, incorporating attention bias with temporal information and replacing standard softmax computation with kernel computation. This not only models time-variant information into an attention score but also reduces both the memory and computation complexity while preserving competitive performance. In addition, we propose a new metric, **MRED_DOH**, to evaluate the difference performance in terms of various listening times in a day, which reflects another critical but unexplored dimension of fairness.

In summary, our contributions are as follows: 1) We propose RecFormer, a novel personalized temporal-aware Transformer for fair music recommendation by adopting masked language modeling tasks as training objectives to learn fine-grained track representations. 2) To reduce the computation complexity and model time-variant information, a temporal-aware linearized-attention is designed by replacing softmax with kernel computation and integrating temporal embeddings into attention scores. Furthermore, we propose a new metric (MRED_DOH) to reflect the different performance of predicting tracks in each hour, which is also an essential but challenging dimension of fairness. 3) Our RecFormer outperforms the official baseline at least 116% in terms of the total score. Moreover, extensive experiments were further conducted to examine the contribution of each module in RecFormer.

2. Related Work

Recommendation System. The recent progress of recommendation systems (RSs) has brought great incomes for industry since the preferred items of the target audience can be marked precisely based on collecting and analyzing their corresponding digital footprints. Early work on RSs typically employed collaborative filtering (CF) [5] and matrix factorization (MF) [6]. Recently, there have been several sequential-based recommendation systems to model user behaviors in the temporal aspect, which was ignored in the early RSs. For instance, Hidasi et al. [7] introduced GRU4Rec with sequential models and ranking loss, which used the idea of taking previous users' records into account to predict future preferences. However, one of the limitations in this task is the validation strategy, which does not guarantee that the test set is from the latest timestamp. This hinders the above approaches to tackle this task effectively. Inspired by the robust pretrained tasks of BERT [8], we aim to adopt masked language modeling tasks to randomly mask the user sequence and reconstruct them, which is able to learn a fine-grained track representation as well as the robustness of the time-invariant test set. This motivation has been used by [9], who proposed BERT4Rec, a bidirectional Transformer with masked language modeling tasks, and demonstrated the effectiveness in several sequential recommendation tasks.

Dataset Introduction. The selected dataset in this task is LFM-1b [10], which is a dataset focused on music recommendation on Last.fm. The dataset is composed of 120k users, 63k artists, 1.3M albums, 821k tracks, and 38M listening events, which is filtered with some preprocessing procedure introduced in [11]. Furthermore, it provides rich song (i.e., artist and album) and user (i.e., country, age, gender, listening preference) metadata for evaluating multi-dimensional behaviors of models. In general, this dataset is able to help researchers achieve not only quantitative performance but also non-standard metrics for fairness.

3. Methodology

Figure 2 illustrates an overview of our proposed RecFormer, some of which is inspired by the recent research on natural language understanding. Given a user history (sequence), we first apply random masks to the sequence, and then personalized user embeddings are generated based on the user metadata and corresponding tracks. Afterwards, the RecFormer, which modifies the self-attention mechanism to the proposed temporal-aware linearized-attention in each layer, is introduced

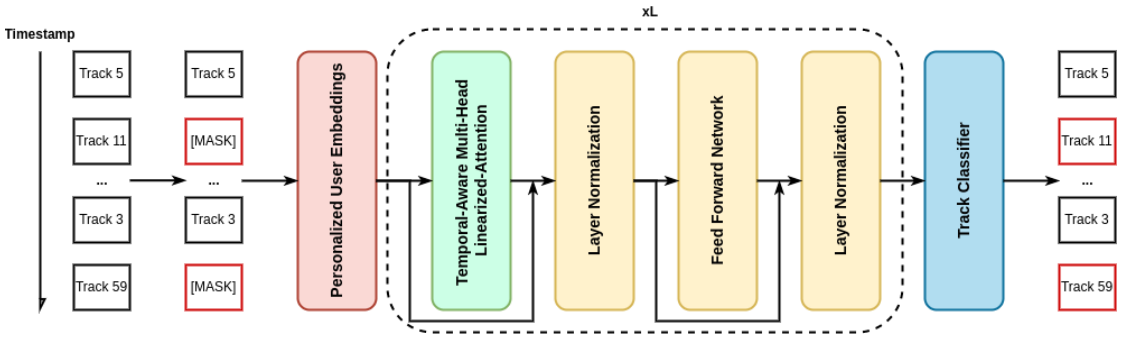

Figure 2: The architecture of our proposed RecFormer.

to model the masked user embeddings. Finally, a track classifier is used to reconstruct the masked tracks to the original tracks.

3.1. Personalized User Embeddings

As each user has multiple tracks they have listened to and the corresponding user metadata, we incorporate each type of user metadata with each track to model each persona, which is inspired by [12]. Specifically, the input embedding for the RecFormer at the i-th timestamp is constructed by adding the track embedding t_i, positional embedding p_i, and metadata embeddings (g, c, a), which are all projected with corresponding embedding layers to d dimensional vectors:

$$E = (e_1, \cdots, e_L) = (t_1 + p_1 + g + c + a, \cdots, t_L + p_L + g + c + a), \tag{1}$$

where g denotes gender, c denotes country, a denotes age and L is the max sequence length. The user age is discretized to 15 bins since it is a continuous variable.

3.2. RecFormer

RecFormer aims to capture the temporal order of a user history, which is hard to consider with traditional CF and MF approaches. To that end, we introduce RecFormer based on the Transformer encoder [4] to not only encode all tracks in a sequence but also to speed up the training procedure with parallel computation of the attention mechanism compared with the recurrent-based models.

Formally, the personalized user embedding E follows the standard Transformer encoding steps to encode the sequence with the proposed temporal-aware linearized-attention (TALA), residual connection and layer normalization (Norm), dropout, and feed forward network (FFN) as follows:

$$\tilde{H} = Norm(E + TALA(E)), H = Norm(\tilde{H} + FFN(\tilde{H})). \tag{2}$$

The output dimension of H is d, and the inner dimension of FFN is d_{inner}.

Temporal-Aware Linearized-Attention (TALA): To reduce the quadratic memory and computation complexity from the conventional self-attention mechanism, we replace the softmax with the kernel computation, which only requires linear computation [13]:

$$Q = EW^Q, K = EW^K, V = EW^V, \tag{3}$$

$$TALA(E) = (\phi(Q)\phi(K)^T)V, \tag{4}$$

where $W^Q, W^K, W^V \in \mathbb{R}^{d \times d}$, and $\phi(\cdot)$ is applied rowwise and is used elu$(\cdot) + 1$ as suggested in [13].

Since users listen to music tracks following their personal habits (e.g., at work or on a bus), it is essential to take time-invariant information into account for recommending user preferred tracks, e.g., [14, 3]. Therefore, in addition to the linearized-attention, we also incorporate the listening time of each track as the attention bias, which is motivated from [15]. Formally, Equ. 4 is extended as:

$$TALA(E) = (\phi(Q)\phi(K + E_H)^T)V, \tag{5}$$

where E_H is the d dimensional hour embedding (from 0 to 23, total 24 categories). It is noted that we empirically add hour embeddings to key matrices from experiments.

3.3. Track Classifier

After L layers of RecFormer to encode multi-hop information, we get the final output H^L for all items of the user sequence. The track classifier is employed to predict the masked tracks as shown in Figure 2. Specifically, we apply a feed-forward layer to generate the i-th output:

$$Z_i = \sigma(H_i^L W^Z); \hat{y}_i = softmax(Z_i), \tag{6}$$

where $W^Z \in \mathbb{R}^{d \times T}$, T is the total number of tracks, and σ is the activation function.

Table 1
Ablation study of our model. U: Personalized user embeddings. T: Temporal-aware computation in TALA. Total score is computed as ((1) + (2) + (3)) / 3 as Phase 2 requires a minimum hit-rate threshold.

	RecFormer (ours)	-U	-T	-U -T
Standard RSs metrics (1)	**0.0093**	0.0071	0.0092	0.0072
Standard metrics on a per-group (2)	**-0.0061**	-0.0077	-0.0158	-0.0099
Behavioral tests (3)	**-0.0213**	-0.1097	-0.0225	-0.1101
MRED_DOH (ours)	-0.0047	**-0.0030**	-0.0033	**-0.0030**
Total Score	**-0.0060**	-0.0368	-0.0096	-0.0376

Training and Testing. As one of the challenges in this task is the validation strategy, it is expected that the test set may not be sampled at the latest timestamp. To tackle this issue, we applied masked language modeling tasks (MLM) as the training objective to learn a robust track representation. The goal of MLM is to reconstruct the masked tracks by giving a user sequence, which enables the model to learn the relation between tracks.

Following [9], we use the final output H with the track classifier to predict the masked tracks, and the loss function is defined as follows:

$$\mathbb{L} = -\sum_{i=1}^{|U|} y_i log(\hat{y}_i), \tag{7}$$

where U is the set of user sequences.

In the inference phase, we empirically set the mask in the last timestamp of a sequence to predict the most possible 100 tracks that the user are likely to listen to. In addition, as we cannot fetch the timestamp in the test set, we set the predicted hour as the hour that the user often listens to music tracks to fit into TALA.

4. Experiments and Analysis

4.1. Experimental Setup

The dimension d was set to 64, the inner dimension of the feed-forward layer was 256, and the number of heads was set to 1. The dropout rate was 0.0, and the max sequence length (L) was 60 due to the time limit, which kept about 25% tracks on average. The batch size was 100, the learning rate was set to 1e-3, the training epochs were set to 50, and the seeds were tested from 42 to 52. It is noted that the number of predicted tracks is based on the train set. That is, we hypothesize that our RecFormer only recommends tracks that have been listened to before. All the training and evaluation phases were conducted on a machine with AMD Ryzen Threadripper 3960X 24-Core Processor, Nvidia GeForce RTX 3090, and 252GB RAM.

The evaluation metrics include different perspectives: standard RSs metrics (HR, and MRR), standard metrics on a per-group or slice basis (gender balance, artist popularity, user country, song popularity, and user history), and behavioral tests (be less wrong, and latent diversity) [11]. The data are pre-processed by filling NaN values of user gender and country with n and UNKNOWN, respectively.
Proposed MRED_DOH Metric: Since one of the challenges we aim to address is the time-variant event, we propose a new metric, **MRED_DOH**, to evaluate the difference performance in terms of various listening times in a day, which reflects another critical but unexplored dimension of fairness. That is, this MRED_DOH enables us to investigate if a model bias to predictions that are from users listening to in specific time slots. In other words, we operationalize this metric as the smaller the difference similar to MRED_Gender proposed in Reclist [16]: the fairer the model towards potential temporal biases.

Specifically, we represent the hour when the user most often listens to select the test set as sub-groups of listeners (i.e., there are total 24 sub-groups), which is represented as the user hour. Afterwards, we can evaluate the MRED score using the existing RecList with the user hour to measure the model performance at each hour in a day. We note that this is a aspect-driven metric, which can be adjusted by monitoring different temporal dimensions based on user needs. For example, the hour when the user most often listens can be easily changed to the least active hours or the average active hours. Moreover, it can also be employed in the sequential recommendation systems by changing the user hours to sequential positions.

4.2. Overall Performance Comparison

Ablation Study. To verify the contribution of each module in RecFormer, we conduct ablative experiments by removing personalized user embeddings, temporal-aware computation, and both. From Table 1, we can observe that removing any one module in RecFormer results in a performance drop in terms of metrics adopted in this task, which testifies to the effective design of RecFormer. However, our RecFormer performs the worst in terms

Table 2
Official performance of RecFormer. The score is normalized with the official baseline and the best score of Phase 1.

Rank	Model	Score	Standard RSs metrics	Standard metrics on a per-group	Behavioral tests	MRED_DOH
7	RecFormer	0.7526	0.0098	-0.0056	-0.0011	-0.0047
-	BERT4Rec [9]	-100.0	0.0016	-0.0030	-0.2729	-0.0014
-	CBOWRecSysBaseline	-1.2122	0.0512	-3.7194	0.4527	-0.0034
Imp. (%)	-	116	-	-	-	

of our proposed MRED_DOH, which indicates that our method still fails to meet the fairness in the temporal aspect. In addition, this result also indicates that considering temporal-awareness is not able to address the temporal fairness, which will be investigated in our future research.

We note that several continuous variables (e.g., novelty_artist related features) are also included in the personalized user embeddings with projecting as in [17], but the training loss cannot converge.

Official Score. Table 2 shows the performance in the formal phase. We also implemented BERT4Rec to compare the performance, which can be viewed as one of our variants. It can be observed that the MRED_DOH performance of BERT4Rec is the best, but the performance of standard RSs metrics fails to meet the requirement (hit-rate > 0.015). One of the reasons is that BERT4Rec is not converged using the same hyper-parameters due to the computational complexity, which is attributed by linear attention in our RecFormer. Therefore, these fairness results cannot be directly compared with the official baseline and our RecFormer since BERT4Rec cannot recommend possible tracks. Our framework achieved 0.1964 of the total score, which outperformed the official baseline by 116%, while it still has some gaps compared to the first prize. Despite the result, our approach still demonstrates that using MLM as the training objective can achieve competitive performance.

5. Conclusion

In this paper, we propose RecFormer incorporating personalized user embeddings and temporal-aware linearized-attention to recommend accurate and fair tracks to users based on their personal listening habits for the EvalRS task. Furthermore, the linearized-attention reduces both computation and memory complexity by making use of kernel computation. The ablation study with the proposed metric demonstrates the effectiveness and fairness of our proposed approach. From the leaderboard score, our method illustrates that using MLM for learning track representations can achieve competitive performance.

References

[1] C. Luo, V. Lakshman, A. Shrivastava, T. Cao, S. Nag, R. Goutam, H. Lu, Y. Song, B. Yin, ROSE: robust caches for amazon product search, in: WWW (Companion Volume), ACM, 2022, pp. 89–93.

[2] S. Mehta, Why is the fairness in recommender systems required?, 2022. URL: https://analyticsindiamag.com/why-is-the-fairness-in-recommender-systems-required/.

[3] Y. Koren, Collaborative filtering with temporal dynamics, in: Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, 2009, pp. 447–456.

[4] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, in: NIPS, 2017, pp. 5998–6008.

[5] B. M. Sarwar, G. Karypis, J. A. Konstan, J. Riedl, Item-based collaborative filtering recommendation algorithms, in: WWW, ACM, 2001, pp. 285–295.

[6] J. Chen, C. Wang, S. Zhou, Q. Shi, J. Chen, Y. Feng, C. Chen, Fast adaptively weighted matrix factorization for recommendation with implicit feedback, in: AAAI, AAAI Press, 2020, pp. 3470–3477.

[7] B. Hidasi, A. Karatzoglou, L. Baltrunas, D. Tikk, Session-based recommendations with recurrent neural networks, in: ICLR (Poster), 2016.

[8] J. Devlin, M. Chang, K. Lee, K. Toutanova, BERT: pre-training of deep bidirectional transformers for language understanding, in: NAACL-HLT (1), Association for Computational Linguistics, 2019, pp. 4171–4186.

[9] F. Sun, J. Liu, J. Wu, C. Pei, X. Lin, W. Ou, P. Jiang, Bert4rec: Sequential recommendation with bidirectional encoder representations from transformer, in: CIKM, ACM, 2019, pp. 1441–1450.

[10] M. Schedl, The lfm-1b dataset for music retrieval and recommendation, in: ICMR, ACM, 2016, pp. 103–110.

[11] J. Tagliabue, F. Bianchi, T. Schnabel, G. Attanasio, C. Greco, G. de Souza P. Moreira, P. J. Chia, Evalrs: a rounded evaluation of recommender systems, CoRR abs/2207.05772 (2022).

[12] Y. Zheng, R. Zhang, M. Huang, X. Mao, A pre-training based personalized dialogue generation model with persona-sparse data, in: AAAI, AAAI Press, 2020, pp. 9693–9700.

[13] A. Katharopoulos, A. Vyas, N. Pappas, F. Fleuret, Transformers are rnns: Fast autoregressive transformers with linear attention, in: ICML, volume 119 of *Proceedings of Machine Learning Research*, PMLR, 2020, pp. 5156–5165.

[14] J. Bao, Y. Zhang, Time-aware recommender system via continuous-time modeling, in: CIKM, 2021, pp. 2872–2876.

[15] P. Shaw, J. Uszkoreit, A. Vaswani, Self-attention with relative position representations, in: NAACL-HLT (2), Association for Computational Linguistics, 2018, pp. 464–468.

[16] P. J. Chia, J. Tagliabue, F. Bianchi, C. He, B. Ko, Beyond NDCG: behavioral testing of recommender systems with reclist, in: WWW (Companion Volume), ACM, 2022, pp. 99–104.

[17] W. Wang, H. Shuai, K. Chang, W. Peng, Shuttlenet: Position-aware fusion of rally progress and player styles for stroke forecasting in badminton, in: AAAI, AAAI Press, 2022, pp. 4219–4227.