Towards Universal Sequence Representation Learning for Recommender Systems

Yupeng Hou[*,†]
houyupeng@ruc.edu.cn
Gaoling School of Artificial Intelligence, Renmin University of China

Shanlei Mu[*]
slmu@ruc.edu.cn
School of Information, Renmin University of China

Wayne Xin Zhao[†,‡,✉]
batmanfly@gmail.com
Gaoling School of Artificial Intelligence, Renmin University of China

Yaliang Li
yaliang.li@alibaba-inc.com
Alibaba Group

Bolin Ding
bolin.ding@alibaba-inc.com
Alibaba Group

Ji-Rong Wen[†]
jrwen@ruc.edu.cn
Gaoling School of Artificial Intelligence, Renmin University of China

ABSTRACT

In order to develop effective sequential recommenders, a series of sequence representation learning (SRL) methods are proposed to model historical user behaviors. Most existing SRL methods rely on explicit item IDs for developing the sequence models to better capture user preference. Though effective to some extent, these methods are difficult to be transferred to new recommendation scenarios, due to the limitation by explicitly modeling item IDs. To tackle this issue, we present a novel universal sequence representation learning approach, named **UniSRec**. The proposed approach utilizes the associated description text of items to learn transferable representations across different recommendation scenarios. For learning *universal item representations*, we design a lightweight item encoding architecture based on parametric whitening and mixture-of-experts enhanced adaptor. For learning *universal sequence representations*, we introduce two contrastive pre-training tasks by sampling multi-domain negatives. With the pre-trained universal sequence representation model, our approach can be effectively transferred to new recommendation domains or platforms in a parameter-efficient way, under either inductive or transductive settings. Extensive experiments conducted on real-world datasets demonstrate the effectiveness of the proposed approach. Especially, our approach also leads to a performance improvement in a cross-platform setting, showing the strong transferability of the proposed universal SRL method. The code and pre-trained model are available at: https://github.com/RUCAIBox/UniSRec.

CCS CONCEPTS

• **Information systems → Recommender systems**.

KEYWORDS

Sequential Recommendation, Universal Representation Learning

ACM Reference Format:
Yupeng Hou, Shanlei Mu, Wayne Xin Zhao, Yaliang Li, Bolin Ding, Ji-Rong Wen. 2022. Towards Universal Sequence Representation Learning for Recommender Systems. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '22), August 14–18, 2022, Washington, DC, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3534678.3539381

1 INTRODUCTION

In the literature of recommender systems, sequential recommendation is a widely studied task [6, 10], aiming to recommend suitable items to a user given her/his historical interaction records. Various methods have been proposed to improve the performance of sequential recommendation, from early matrix factorization (*e.g.*, FPMC [18]) to recent sequence neural networks (*e.g.*, GRU4Rec [6], Caser [24] and Transformer [10, 36]). These approaches have largely raised the performance bar of sequential recommendation.

Though the adopted techniques are different, the core idea of existing methods is similar: it first formulates the user behavior as a chronologically-ordered interaction sequence with the items, and then develops effective architectures for capturing the sequential interaction characteristics that reflect the user preference. In this way, the learned sequence models can predict the likely items to be interacted by a user, given the observed sequential context. Framed in the paradigm of representation learning [1], such an approach essentially aims to conduct a sequence representation learning (SRL) model based on historical user behavior data, which needs to effectively capture the item characteristics and sequential interaction characteristics. The capacity of the designed SRL model directly affects the performance of sequential recommendation.

Despite the progress, most of existing SRL methods for recommendation rely on explicit *item IDs* for developing the sequence models [6, 10]. A major issue with this modeling way is that the learned model is difficult to be transferred to new recommendation scenarios, even when the underlying data forms are exactly

[*] Equal contribution.
[†] Beijing Key Laboratory of Big Data Management and Analysis Methods.
[‡] Beijing Academy of Artificial Intelligence, Beijing, 100084, China.
[✉] Corresponding author.

KDD '22, August 14–18, 2022, Washington, DC, USA.
© 2022 Association for Computing Machinery.
ACM ISBN 978-1-4503-9385-0/22/08…$15.00
https://doi.org/10.1145/3534678.3539381

the same. Such an issue limits the reuse of the recommendation model across domains [39]. As a common case, we need to re-train a sequential recommender from scratch when adapting to a new domain, which is tedious and resource-consuming. In addition, existing sequential recommenders usually suffer from the recommendation of cold-start items having inadequate interactions with users. For addressing the above issue, a number of studies have been proposed by learning either semantic mapping [38] or transferable components [12], in order to bridge the domain gap and enhance the item representations. However, these existing attempts cannot fully solve the fundamental issue caused by explicitly modeling item IDs. More recently, increasing evidence shows that natural language text can play the role of general semantic bridge across different tasks or domains [3] (including the recommendation tasks, *e.g.,* zero-shot recommendation [4]), with the remarkable success of pre-trained language models (PLM).

Inspired by the recent progress of language intelligence [3, 26], we aim to design a new SRL approach for learning more generalizable sequence representations, by breaking the limit of explicit ID modeling. The core idea is to utilize the associated description text (*e.g.,* product description, product title or brand) of an item, called *item text*, to learn transferable representations across different domains. Although previous attempts have shown such an approach is promising [4], there are still two major challenges to be solved. First, the textual semantic space is not directly suited for the recommendation tasks. It is not clear how to model and utilize item texts for improving the recommendation performance, since directly introducing raw textual representations as additional features may lead to suboptimal results. Second, it is difficult to leverage multi-domain data for improving the target domain, where the *seesaw phenomenon* (referring to learning from multiple kinds of domain-specific patterns is conflict or oscillating) often appears [23].

To address the above issues, we propose the universal sequence representation learning approach, named **UniSRec**. Our approach takes general interaction sequences as input, and learns universal ID-agnostic representations based on a pre-training approach. Specially, we focus on two key points that learn *universal item representation* and *universal sequence representations.* For learning universal item representations, we design a lightweight architecture based on parametric whitening and mixture-of-experts enhanced adaptor, which can derive more isotropic semantic representations as well as enhance the domain fusion and adaptation. For learning universal sequence representations, we introduce two kinds of contrastive learning tasks, namely sequence-item and sequence-sequence contrastive tasks, by sampling multi-domain negatives. Based on the above methods, the pre-trained model can be effectively transferred to a new recommendation scenario in a parameter-efficient way, under either inductive or transductive settings.

To evaluate the proposed approach UniSRec, we conduct extensive experiments on real-world datasets from different application domains and platforms. Experimental results demonstrate that the proposed approach can effectively utilize data from multiple domains to learn universal and transferable representations. Especially, the results on cross-platform experiments show that recommendation performance can be improved with the universal sequence representation model pre-trained on other platforms without overlapping users or items.

2 METHODOLOGY

In this section, we present the proposed **Uni**versal **S**equence representation learning approach for **Rec**ommendation, named as **UniSRec**. Given historical user behavioral sequences from a mixture of multiple domains, we aim to learn universal item and sequence representations that can effectively transfer and generalize to new recommendation scenarios (*e.g.,* new domains or platforms) in a parameter-efficient way.

2.1 Overview of the Approach

Our approach takes general interaction sequences as input and learns universal representations based on a pre-training approach. We then formulate the task and overview the proposed approach.

General input formulation. We formulate the behavior sequence of a user in a general form of interaction sequence $s = \{i_1, i_2, \cdots i_n\}$, (in a chronological order of interaction time), where each interacted item i is associated with a unique item ID and a description text (*e.g.,* the product description, item title or brand). We call the description text of an item i *item text*, denoted by $t_i = \{w_1, w_2, \cdots, w_c\}$, where the words w_j are from a shared vocabulary and c denotes the truncated length of item text. Here, each sequence contains all the interaction behavior of a user at some specific domain, and a user can generate multiple behavior sequences at different domains or platforms. As discussed before, as there are large semantic gap between different domains, we don't simply mix the behavior data of a user. Instead, we treat the multiple interaction sequences of a user as different sequences, without explicitly maintaining user IDs for each sequence. Note that unlike other pre-training based recommendation methods [31, 36], item IDs are only auxiliary information in our approach, and we mainly utilize item text to derive generalizable ID-agnostic representations. Unless specified, item IDs will not be used as input of our approach.

我们主要利用项目文本来推导
可泛化的与 ID 无关的表示

Solutions. To learn transferable representations across domains, we identify two key problems for achieving this purpose, *i.e.,* learning universal item representation and sequence representation, since *items* and *sequences* are the basic data forms in our general formulation. For learning universal item representations (Section 2.2), we focus on the domain fusion and adaptation with an MoE-enhanced adaptor based on parametric whitening. For learning universal sequence representations (Section 2.3), we introduce two kinds of contrastive learning tasks, namely sequence-item and sequence-sequence contrastive tasks, by sampling multi-domain negatives. Based on the above methods, the pre-trained model can be effectively transferred to a new recommendation scenario in a parameter-efficient way, under either inductive or transductive settings (Section 2.4). The overall framework of the proposed approach UniSRec is depicted in Figure 1.

2.2 Universal Textual Item Representation

The first step toward universal sequential behavior modeling is to represent items from various recommendation scenarios (*e.g.,* domains or platforms) into a unified semantic space. In previous studies [6, 10], item representations are usually learned in *transductive learning* setting, where item IDs are pre-given and ID embeddings are learned as item representations. Such a way largely limits the

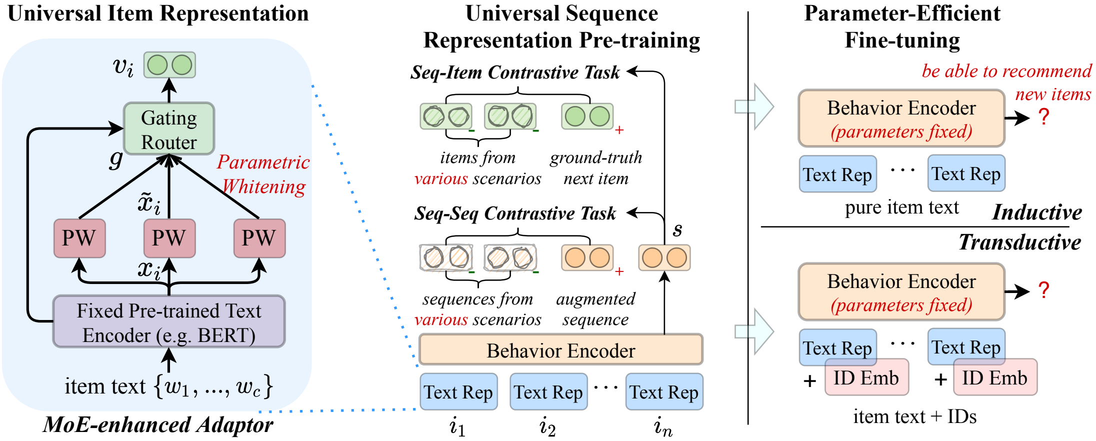
Universal Item Representation
v_i
Gating Router
g
Parametric Whitening
$\tilde{x}_i$
PW
PW
PW
x_i
Fixed Pre-trained Text Encoder (e.g. BERT)
item text $\{w_1, ..., w_c\}$
MoE-enhanced Adaptor
Universal Sequence Representation Pre-training
Seq-Item Contrastive Task
items from various scenarios
ground-truth next item
Seq-Seq Contrastive Task
s
sequences from various scenarios
augmented sequence
Behavior Encoder
Text Rep
Text Rep
Text Rep
i_1
i_2
i_n
Parameter-Efficient Fine-tuning
be able to recommend new items
Behavior Encoder (parameters fixed)
?
Text Rep
Text Rep
pure item text
Inductive
Transductive
Behavior Encoder (parameters fixed)
?
Text Rep
Text Rep
+ ID Emb
+ ID Emb
item text + IDs

Figure 1: The overall framework of the proposed universal sequence representation learning approach (UniSRec).

transferability of item representations, since the vocabularies of item IDs are usually different across domains.

Our solution is to learn transferable item representations based on the associated item text, which describes the item characteristics in the form of natural language. As more and more evidence shows [3], natural language provides a general data form to bridge the semantic gap across different tasks or domains. Based on this idea, we first utilize the pre-trained language model (PLMs) to learn the text embeddings. Since the derived text representations from different domains are likely to span different semantic spaces (even with the same text encoder), we propose the techniques of parametric whitening and mixture-of-experts (MoE) enhanced adaptor to transform the text semantics into a universal form suited to the recommendation tasks.

2.2.1 Textual Item Encoding via Pre-trained Language Model. Considering the excellent language modeling capacity of PLMs, we utilize the widely used BERT model [3] to learn universal text representations for representing items. Given an item i and its corresponding text t_i, we firstly concatenate (1) a special symbol [CLS], (2) the words of item text $\{w_1, w_2, \ldots, w_c\}$, in order and derive the input sequence for BERT. Then we feed the concatenated sequence into the BERT model, and we have:

$$\boldsymbol{x}_i = \text{BERT}([\texttt{[CLS]}; w_1, \ldots, w_c]), \tag{1}$$

where $\boldsymbol{x}_i \in \mathbb{R}^{d_W}$ is the final hidden vector corresponding to the first input token ([CLS]), and "[;]" denotes the concatenation operation.

2.2.2 Semantic Transformation via Parametric Whitening. Though we can obtain semantic representations from BERT, they are not directly suited for the recommendation tasks. Existing studies [11] have found that BERT induces a non-smooth anisotropic semantic space for general texts. The case will become more severe when we mix item texts from multiple domains with a large semantic gap. Inspired by recent works on whitening-based methods [9, 21], we conduct a simple linear transformation to transform original BERT representations for deriving isotropic semantic representations.

BERT 为一般文本引入了非平滑的各向异性语义空间

Different from the original whitening methods [9, 21] with preset mean and variance, we incorporate learnable parameters in the whitening transformation for better generalizability on unseen domains. Formally, we have:

$$\widetilde{\boldsymbol{x}}_i = (\boldsymbol{x}_i - \boldsymbol{b}) \cdot \boldsymbol{W}_1, \tag{2}$$

where $\boldsymbol{b} \in \mathbb{R}^{d_W}$ and $\boldsymbol{W}_1 \in \mathbb{R}^{d_W \times d_V}$ are learnable parameters, and $\widetilde{\boldsymbol{x}}_i \in \mathbb{R}^{d_V}$ is the transformed representation. In this way, the anisotropy issue of learned representations can be alleviated, which is useful to learn universal semantic representations. For the efficiency consideration, we don't introduce complex non-linear architectures, such as flow-based generative models [11], which will be studied in future work.

2.2.3 Domain Fusion and Adaptation via MoE-enhanced Adaptor. With the above whitening transformation, our model can learn more isotropic 向同性 semantic representations. In order to learn universal item representations, another important issue is how to transfer and fuse information across domains, since there is usually a large semantic gap between different domains. For example,the top frequent words of item text are quite different across domains, *e.g.*, *natural*, *sweet*, *fresh* for food domain and *war*, *love*, *story* for movies domain. A straightforward approach is to map the original BERT embeddings into some shared semantic space [11]. However, it will lead to limited representation capacity for migrating the domain bias. As our solution, we learn multiple whitening embeddings for an item, and utilize an adaptive combination of these embeddings as the universal item representations. Not limited to a simple single mapping between item representations, we aim to establish a more flexible representation mechanism to capture the semantic relatedness for domain fusion and adaptation.

To implement our idea, we employ the mixture-of-expert (MoE) architecture [19] for learning more generalizable item representations. Specially, we incorporate G whitening transformation modules as the *experts*, and then construct the MoE-enhanced adaptor based on a parameterized router:

$$\boldsymbol{v}_i = \sum_{k=1}^{G} g_k \cdot \widetilde{\boldsymbol{x}}_i^{(k)}, \tag{3}$$

where $\widetilde{\boldsymbol{x}}_i^{(k)}$ is the output of the k-th whitening transformation module (Eqn. (2)) and g_k is the corresponding combination weight from the *gating router*, defined as follows:

$$\boldsymbol{g} = \text{Softmax}\left(\boldsymbol{x}_i \cdot \boldsymbol{W}_2 + \boldsymbol{\delta}\right), \tag{4}$$

$$\boldsymbol{\delta} = \text{Norm}() \cdot \text{Softplus}\left(\boldsymbol{x}_i \cdot \boldsymbol{W}_3\right). \tag{5}$$

In this equation, we utilize original BERT embedding $\boldsymbol{x}_i$ as input of the router module, as it contains domain-specific semantic bias. Furthermore, we incorporate learnable parameter matrices $\boldsymbol{W}_2, \boldsymbol{W}_3 \in \mathbb{R}^{d_W \times G}$ to adaptively adjust the weights of experts $\boldsymbol{g} \in \mathbb{R}^G$. In order to balance the expert load, we utilize Norm() to generate random Gaussian noise, controlled by parameter $\boldsymbol{W}_2$.

MoE 强化适配器的优点有三重

The merits of the MoE-enhanced adaptor are threefold. Firstly, the representation of a single item is enhanced by learning multiple whitening transformations. Second, we no longer require a direct semantic mapping across domains, but instead utilize a learnable gating mechanism to adaptively establish the semantic relatedness for domain fusion and adaptation. Third, the lightweight adaptor endows the flexibility of parameter-efficient fine-tuning when adapting to new domains (detailed in Section 2.4).

2.3 Universal Sequence Representation

Since different domains usually correspond to varying user behavioral patterns, it may not work well to simply mix interaction sequences from multiple domains for pre-training. And, it is likely to lead to the *seesaw phenomenon* [23] that the learning from multiple domain-specific behavioral patterns can be conflict. Our solution is to introduce two kinds of contrastive learning tasks, which can further enhance the fusion and adaptation of different domains in deriving the item representations. In what follows, we firstly introduce the base behavior encoder architecture, and then present the proposed contrastive pre-training tasks that enhance the sequence representations in universal semantic space.

2.3.1 *Self-attentive Sequence Encoding.*

Given a sequence of universal item representations, we further utilize a user behavior encoder to obtain the sequence representation. We aim to construct the sequential patterns based on the learned universal textual item representations, but not item IDs. Here, we adopt a widely used self-attentive architecture [26], *i.e.,* Transformers. Specially, it consists stacks of multi-head self-attention layers (denoted by MHAttn(·)) and point-wise feed-forward networks (multilayer perceptron activated by ReLU, denoted by FFN(·)). We sum the learned text representations (*i.e.,* $\boldsymbol{v}_i$ in Eqn. (3)) and the absolute position embeddings $\boldsymbol{p}_j$ as input at position j. The input and the update process can be formalized as following:

$$\boldsymbol{f}_j^0 = \boldsymbol{v}_i + \boldsymbol{p}_j, \tag{6}$$

$$\boldsymbol{F}^{l+1} = \text{FFN}(\text{MHAttn}(\boldsymbol{F}^l)), \tag{7}$$

where $\boldsymbol{F}^l = [\boldsymbol{f}_0^l; \dots; \boldsymbol{f}_n^l]$ denotes the concatenated representations at each position in the l-th layer. We take the final hidden vector $\boldsymbol{f}_n^L$ corresponding to the n-th (last) position as the sequence representation (a total number of L layers in the behavior encoder).

2.3.2 *Multi-domain Sequential Representation Pre-training.*

Given interaction sequences from multiple domains, we next study how to design suitable optimization objectives to derive the outputs of sequential encoder in the unified representation space. By contrasting sequences and items from different domains, we aim to alleviate the seesaw phenomenon and capture their semantic correlation in the pre-training stage. For this purpose, we design the following sequence-item and sequence-sequence contrastive tasks.

Sequence-item contrastive task. The sequence-item contrastive task aims to capture the intrinsic correlation between sequential contexts (*i.e.,* the observed subsequence) and potential next items in an interaction sequence. Different from previous next-item prediction task [10, 15] (using *in-domain* negatives), for a given sequence, we adopt *across-domain* items as negatives. Such a way can enhance both the semantic fusion and adaptation across domains, which is helpful to learn universal sequence representations.

We consider a batch setting of B training instances, where each training instance is a pair of the sequential context (containing the proceeding items) and the positive next item. We first encode them into embedding representations $\{\langle \boldsymbol{s}_1, \boldsymbol{v}_1 \rangle, \dots, \langle \boldsymbol{s}_B, \boldsymbol{v}_B \rangle\}$, where $\boldsymbol{s}$ represents the normalized contextual sequence representations, and $\boldsymbol{v}$ denotes the representation of the positive next item. Then, we formalize the sequence-item contrastive loss as follows:

$$\ell_{S-I} = -\sum_{j=1}^{B} \log \frac{\exp\left(\boldsymbol{s}_j \cdot \boldsymbol{v}_j / \tau\right)}{\sum_{j'=1}^{B} \exp\left(\boldsymbol{s}_j \cdot \boldsymbol{v}_{j'} / \tau\right)}, \tag{8}$$

where in-batch items are regarded as negative instances and τ is a temperature parameter. As batches are constructed randomly, the in-batch negative instances $\{\boldsymbol{v}_{j'}\}$ will contain items from a mixture of multiple domains.

Sequence-sequence contrastive task. Besides the above item-level pre-training task, we further propose a sequence-level pre-training task, by conducting the contrastive learning among multi-domain interaction sequences. The object is to discriminate the representations of augmented sequences from multi-domain sequences. We consider two kinds of augmentation strategies: (1) *Item drop* refers to randomly dropping a fixed ratio of items in the original sequence, and (2) *Word drop* refers to randomly dropping words in item text. Given a target sequence (with the representation $\boldsymbol{s}_j$), the augmented ones are considered as positives (with the representation $\widetilde{\boldsymbol{s}}_j$), while other in-batch ones are considered as negatives. The sequence-sequence contrastive loss can be formally presented as following:

$$\ell_{S-S} = -\sum_{j=1}^{B} \log \frac{\exp\left(\boldsymbol{s}_j \cdot \widetilde{\boldsymbol{s}}_j / \tau\right)}{\sum_{j'=1}^{B} \exp\left(\boldsymbol{s}_j \cdot \boldsymbol{s}_{j'} / \tau\right)}. \tag{9}$$

Similar to Eqn. (8), as batches are constructed randomly, the in-batch negative instances naturally contain sequences from multiple domains. In the implementation, we preprocess the augmented item text using *word drop* for efficient pre-training, as the BERT representations of item text can be obtained during preprocessing.

Multi-task learning. At the pre-training stage, we leverage a multi-task training strategy to jointly optimize the proposed sequence-item contrastive loss in Eqn. (8) and sequence-sequence contrastive loss in Eqn. (9):

$$\mathcal{L}_{\text{PT}} = \ell_{S-I} + \lambda \cdot \ell_{S-S}, \tag{10}$$

where λ is a hyper-parameter to control the weight of sequence-sequence contrastive loss. The pre-trained model is to be fine-tuned for adapting to new domains.

2.4 Parameter-Efficient Fine-tuning

To adapt to a new domain, previous pre-training based recommendation methods [4, 36] usually require fine-tuning the whole network architecture, which is time-consuming and less flexible. Since our model can learn universal representations for interaction sequences, our idea is to fix the parameters of the major architecture, while only fine-tuning a small proportion of parameters from the MoE-enhanced adaptor (Section 2.2.3) for incorporating necessary adaptation. We find that the proposed MoE-enhanced adaptor can quickly adapt to unseen domains, fusing the pre-trained model with new domain characteristics. To be specific, we consider two fine-tuning settings, either *inductive* or *transductive*, based on whether item IDs in the target domain can be accessed.

Inductive setting. The first setting considers the test cases of recommending new items from an unseen domain, which can't be well solved by an ID-based recommendation model. The proposed model doesn't rely on item IDs, so that it can learn universal text representations for new items. Given a training sequence from the target domain, we firstly encode the sequential context $(i_1 \rightarrow i_t)$ and the candidate item i_{t+1} into universal representations as $\bm{s}$ and $\bm{v}_{i_{t+1}}$. Then we predict the next item according to the following probability:

$$P_I(i_{t+1}|s) = \text{Softmax}(\bm{s} \cdot \bm{v}_{i_{t+1}}), \tag{11}$$

where we compute the softmax probability over the candidate set (the positive item and a number of sampled negatives). The parameters to be tuned are those in $\bm{b}$ and $\bm{W}_r$ in Eqn. (2).

Transductive setting. The second setting assumes that nearly all the items of the target domain have appeared in the training set, and we can also learn ID embeddings since item IDs are available. In this setting, to represent an item, we combine the textual embedding $\bm{v}_{i_{t+1}}$ and ID embedding $\bm{e}_{i_{t+1}}$ as the final item representation. Thus, we have the following prediction probability:

$$P_T(i_{t+1}|s) = \text{Softmax}\left(\widetilde{\bm{s}} \cdot (\bm{v}_{i_{t+1}} + \bm{e}_{i_{t+1}})\right), \tag{12}$$

where $\widetilde{\bm{s}}$ denotes the enhanced universal sequence representation by adding ID embeddings in the input (Eqn. (7)). Note that the rest parameters of the sequence encoder are still fixed in this setting.

For each setting, we optimize the widely used cross-entropy loss to fine-tune parameters of MoE-enhanced adaptors. After fine-tuning, we predict the probability distribution of next item for given sequences via Eqn. (11) and Eqn. (12).

2.5 Discussion

In the literature of recommender systems, a large number of recommendation models have been developed. Here, we make a brief comparison with related recommendation models, in order to highlight the novelty and differences of our approach.

General sequential approaches such as GRU4Rec [6] and SASRec [10] rely on explicit item IDs to construct the sequential model, where they assume item IDs are pre-given. These approaches can't perform well under the cold-start setting with new items. As a comparison, our approach aims to construct an ID-agnostic recommendation model that can capture sequential patterns in a more general form of natural language.

Table 1: Comparison of the transfer learning scenarios and application settings of several approaches. $1 \rightarrow 1$ denotes 1 source domain to 1 target domain, and $M \rightarrow N$ denotes M source domains to N target domains. "Non-OL" denotes that the approach doesn't require overlapped users or items.

Methods	Transfer Learning Scenarios			Application Settings	
	$1 \rightarrow 1$	$M \rightarrow N$	Non-OL	Transductive	Inductive
S^3-Rec [36]	✗	✗	✗	✓	✗
PeterRec [31]	✓	✗	✗	✓	✗
RecGURU [12]	✓	✗	✓	✓	✗
ZESRec [4]	✓	✗	✓	✗	✓
UniSRec (ours)	✓	✓	✓	✓	✓

有疑问

Cross-domain approaches such as RecGURU [12] propose to leverage the auxiliary information from source domains to improve the performance on a target domain. However, most approaches require overlapping users or items as anchors. Besides, it is not easy to transfer and fuse multiple source domains for improving the target domain. For our approach, we propose an MoE-enhanced adaptor mechanism for domain fusion and adaptation based on universal textual semantics.

Pre-training sequential approaches mainly pre-train their model on sequences from (1) the current domain (S^3-Rec [36], IDA-SR [16]), (2) other domains with overlapping users (PeterRec [31]) or (3) other closely related domains (ZESRec [4]). However, none of these methods explore how to pre-train universal item and sequence representations on multiple weakly related or irrelevant domains. In contrast, our approach can learn more transferable representations that well generalize to the target domains, from a mixture of multiple source domains. The comparison of these approaches is presented in Table 1.

3 EXPERIMENTS

In this section, we first set up the experiments, and then present the results and analysis.

3.1 Experimental Setup

3.1.1 Datasets. To evaluate the performance of the proposed approach, we conduct experiments in both cross-domain setting and cross-platform setting. The statistics of datasets after preprocessing are summarized in Table 2.

(1) **Pre-trained datasets**: we select five categories from Amazon review datasets [17], "*Grocery and Gourmet Food*", "*Home and Kitchen*", "*CDs and Vinyl*", "*Kindle Store*" and "*Movies and TV*", as the source domain datasets for pre-training.

(2) **Cross-domain datasets**: we select another five categories from Amazon review datasets [17], "*Prime Pantry*", "*Industrial and Scientific*", "*Musical Instruments*", "*Arts, Crafts and Sewing*" and "*Office Products*", as target domain datasets to evaluate the proposed approach in cross-domain setting.

Table 2: Statistics of the datasets after preprocessing. "Avg. n" denotes the average length of item sequences. "Avg. c" denotes the average number of tokens in item text.

Datasets	#Users	#Items	#Inters.	Avg. n	Avg. c
Pre-trained	1,361,408	446,975	14,029,229	13.51	139.34
- Food	115,349	39,670	1,027,413	8.91	153.40
- CDs	94,010	64,439	1,118,563	12.64	80.43
- Kindle	138,436	98,111	2,204,596	15.93	141.70
- Movies	281,700	59.203	3,226,731	11.45	97.54
- Home	731,913	185,552	6,451,926	8.82	168.89
Scientific	8,442	4,385	59,427	7.04	182.87
Pantry	13,101	4,898	126,962	9.69	83.17
Instruments	24,962	9,964	208,926	8.37	165.18
Arts	45,486	21,019	395,150	8.69	155.57
Office	87,436	25,986	684,837	7.84	193.22
Online Retail	16,520	3,469	519,906	26.90	27.80

(3) **Cross-platform datasets**: we also select a dataset from different platforms to evaluate the pre-trained universal sequence representation model in a cross-platform setting. **Online Retail**[1] contains transactions occurring between 01/12/2010 and 09/12/2011 from a UK-based online retail platform, which does not contain shared users or items with the Amazon platform英国在线零售平台

Following previous works [10, 36], we keep the five-core datasets and filter users and items with fewer than five interactions for all datasets. Then we group the interactions by users and sort them by timestamp ascendingly. For item text, we concatenate fields including *title*, *categories* and *brand* in Amazon dataset and directly use the *Description* field in Online Retail dataset. We truncate item text longer than 512 tokens.

3.1.2 Compared Methods. We compare the proposed approach with the following baseline methods:

- **SASRec** [10] adopts a self-attention network to capture the user's preference within a sequence.
- **BERT4Rec** [22] adapts the original text-based BERT model with the cloze objective for modeling user behavior sequences.
- **FDSA** [32] proposes to capture item and feature transition patterns via self-attentive networks.
- **S^3-Rec** [36] pre-trains sequential models via mutual information maximization objectives for feature fusion.
- **CCDR** [28] proposes intra-domain and inter-domain contrastive objects for cross-domain recommendation in matching. We extract textual tags using TF-IDF algorithm, and then optimize the taxonomy-based inter-CL objective.
- **RecGURU** [12] proposes to pre-train user representations via autoencoder in an adversarial learning paradigm. In our implementation, we remove constraints on overlapped users.
- **ZESRec** [4] encodes item text via pre-trained language model as item representations. ZESRec can be pre-trained on source domain and directly applied to target domains for zero-shot recommendation. For a fair comparison, we fine-tune the pre-trained model using item sequences of the target domain.

For our approach, we firstly pre-train a universal sequence representation model on the five source datasets that are introduced in Section 3.1.1. We consider two major variants: (1) **UniSRec$_t$** denotes the model fine-tuned in inductive setting using only item text; (2) **UniSRec$_{t+ID}$** denotes the model fine-tuned in transductive setting using both item ID and item text.

3.1.3 Evaluation Settings. To evaluate the performance of the next item prediction task, we adopt two widely used metrics Recall@N and NDCG@N, where N is set to 10 and 50. Following previous works [22, 33, 36], we apply the leave-one-out strategy for evaluation. For each user interaction sequence, the last item is used as the test data, the item before the last one is used as the validation data, and the remaining interaction records are used for training. We rank the ground-truth item of each sequence among all the other items for evaluation on test set, and finally report the average score of all test users.

3.1.4 Implementation Details. We implement UniSRec using a popular open-source recommendation library RecBole[2] [35]. To ensure a fair comparison, we optimize all the methods with Adam optimizer and carefully search the hyper-parameters of all the compared methods. The batch size is set to 2,048. We adopt early stopping with the patience of 10 epochs to prevent overfitting, and NDCG@10 is set as the indicator. We tune the learning rate in $\{0.0003, 0.001, 0.003, 0.01\}$ and the embedding dimension in $\{64, 128, 300\}$. We pre-train the proposed approach for 300 epochs with $\lambda = 1e^{-3}$ and $G = 8$ experts.

3.2 Overall Performance

We compare the proposed approach with the baseline methods on the five cross-domain datasets and one cross-platform dataset. Note that, we fine-tune the same pre-trained universal sequence representation model on these six datasets for our approach. The results are reported in Table 3.

For the baseline methods, text-enhanced sequential recommendation methods (*i.e.,* FDSA and S^3-Rec) perform better than the traditional sequential recommendation methods (*i.e.,* SASRec and BERT4Rec) on several datasets, since item texts are used as auxiliary features to improve the performance. The cross-domain methods CCDR and RecGURU do not perform well, because they are still in a transductive setting. These methods become less effective without explicit overlapping users between source and target domains. ZESRec utilizes PLMs to encode item texts, but mainly reuses existing architectures and pre-training tasks, which can't fully leverage multi-domain interaction data for improving the target domain.

Finally, by comparing the proposed approach UniSRec$_{t+ID}$ with all the baselines, it is clear that UniSRec$_{t+ID}$ achieves the best performance in almost all the cases. Different from these baselines, we derive universal sequence representations via pre-training on multi-domain datasets. With the specially designed parametric whitening module and MoE-enhanced adaptor module, the learned universal item representations are more isotropic and suitable for domain fusion and adaptation. Especially, the results on cross-platform evaluation (*i.e.,* Online Retail dataset) show that our approach can be effectively transferred to a different platform via universal sequence representation pre-training. Besides, the model fine-tuned

[1] https://www.kaggle.com/carrie1/ecommerce-data

[2] https://recbole.io

Table 3: Performance comparison of different recommendation models. The best and the second-best performances are denoted in bold and underlined fonts, respectively. "Improv." indicates the relative improvement ratios of the proposed approach over the best performance baselines. "*" denotes that the improvements are significant at the level of 0.01 with paired t-test.

Scenario	Dataset	Metric	SASRec	BERT4Rec	FDSA	S^3-Rec	CCDR	RecGURU	ZESRec	UniSRec$_t$	UniSRec$_{t+ID}$	Improv.
Cross-Domain	Scientific	Recall@10	0.1080	0.0488	0.0899	0.0525	0.0695	0.1023	0.0851	0.1188*	**0.1235***	+14.35%
		NDCG@10	0.0553	0.0243	0.0580	0.0275	0.0340	0.0572	0.0475	**0.0641***	0.0634*	+10.52%
		Recall@50	0.2042	0.1185	0.1732	0.1418	0.1647	0.2022	0.1746	0.2394*	**0.2473***	+21.11%
		NDCG@50	0.0760	0.0393	0.0759	0.0468	0.0546	0.0786	0.0670	0.0903*	**0.0904***	+15.01%
	Pantry	Recall@10	0.0501	0.0308	0.0395	0.0444	0.0408	0.0469	0.0454	0.0636*	**0.0693***	+38.32%
		NDCG@10	0.0218	0.0152	0.0209	0.0214	0.0203	0.0209	0.0230	0.0306*	**0.0311***	+35.22%
		Recall@50	0.1322	0.1030	0.1151	0.1315	0.1262	0.1269	0.1141	0.1658*	**0.1827***	+38.20%
		NDCG@50	0.0394	0.0305	0.0370	0.0400	0.0385	0.0379	0.0378	0.0527*	**0.0556***	+39.00%
	Instruments	Recall@10	0.1118	0.0813	0.1070	0.1056	0.0848	0.1113	0.0783	0.1189*	**0.1267***	+13.33%
		NDCG@10	0.0612	0.0620	**0.0796**	0.0713	0.0451	0.0681	0.0497	0.0680	0.0748*	–
		Recall@50	0.2106	0.1454	0.1890	0.1927	0.1753	0.2068	0.1387	0.2255*	**0.2387***	+13.34%
		NDCG@50	0.0826	0.0756	0.0972	0.0901	0.0647	0.0887	0.0627	0.0912	**0.0991***	+1.95%
	Arts	Recall@10	0.1108	0.0722	0.1002	0.1003	0.0671	0.1084	0.0664	0.1066	**0.1239***	+11.82%
		NDCG@10	0.0587	0.0479	**0.0714**	0.0601	0.0348	0.0651	0.0375	0.0586	0.0712	–
		Recall@50	0.2030	0.1367	0.1779	0.1888	0.1478	0.1979	0.1323	0.2049*	**0.2347***	+15.62%
		NDCG@50	0.0788	0.0619	0.0883	0.0793	0.0523	0.0845	0.0518	0.0799	**0.0955***	+8.15%
	Office	Recall@10	0.1056	0.0825	0.1118	0.1030	0.0549	0.1145	0.0641	0.1013	**0.1280***	+11.79%
		NDCG@10	0.0710	0.0634	**0.0868**	0.0653	0.0290	0.0768	0.0391	0.0619	0.0831	–
		Recall@50	0.1627	0.1227	0.1665	0.1613	0.1095	0.1757	0.1113	0.1702	**0.2016***	+14.74%
		NDCG@50	0.0835	0.0721	0.0987	0.0780	0.0409	0.0901	0.0493	0.0769	**0.0991**	+0.41%
Cross-Platform	Online Retail	Recall@10	0.1460	0.1349	0.1490	0.1418	0.1347	0.1467	0.1103	0.1449	**0.1537***	+3.15%
		NDCG@10	0.0675	0.0653	0.0719	0.0654	0.0620	0.0658	0.0535	0.0677	**0.0724**	+0.70%
		Recall@50	0.3872	0.3540	0.3802	0.3702	0.3587	**0.3885**	0.2750	0.3604	**0.3885**	0.00%
		NDCG@50	0.1201	0.1131	0.1223	0.1154	0.1108	0.1188	0.0896	0.1149	**0.1239***	+1.31%

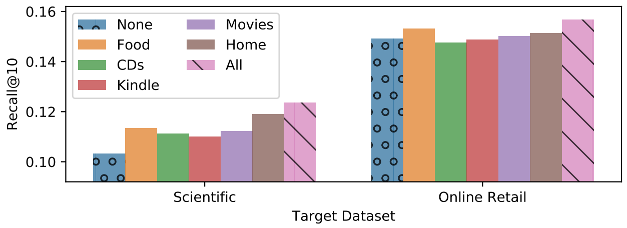

Figure 2: Performance comparison w.r.t. different pre-training datasets on "Scientific" and "Online Retail". "All" denotes the model pre-trained on all five datasets, and "None" denotes the training from scratch.

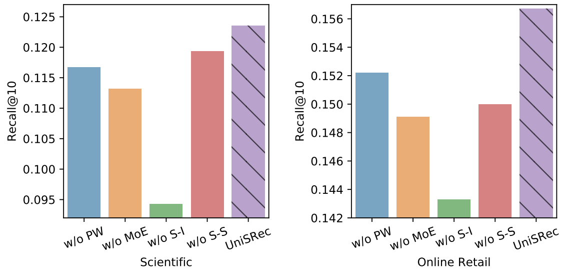

Figure 3: Ablation study of UniSRec variants on "Scientific" and "Online Retail".

in inductive setting (without item IDs, UniSRec$_t$) also has a comparable performance with other baselines. This further illustrates the effectiveness of the proposed universal SRL approach.

3.3 Further Analysis

3.3.1 Universal Pre-training Analysis. In this part, we compare and analyze the effectiveness of universal pre-training. Specifically, we would like to examine whether the universal model pre-trained on multiple datasets performs better than those models pre-trained on one single source dataset, or those models without pre-training. The experimental results are reported in Figure 2.

We can see that the model pre-trained on all the five datasets achieves better performance than any model that pre-trained on a single dataset or without pre-training. The results show that with the proposed universal sequence representation learning approach, the pre-trained model can capture semantic sequential patterns from multiple source domains to improve recommendation on target domains or platforms.

3.3.2 Ablation Study. In this part, we analyze how each of the proposed techniques or components affects the final performance. We prepare four variants of the proposed UniSRec model for comparisons, including (1) *w/o* PW that replaces parametric whitening

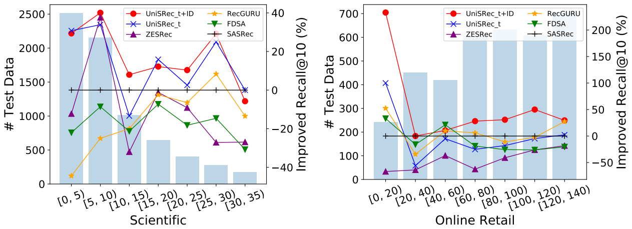

Figure 4: Performance comparison w.r.t. long-tail items on the “Scientific” and “Online Retail” datasets. The bar graph represents the number of interactions in test data for each group. The line chart represents the improvement ratios for Recall@10 compared with SASRec.

with traditional linear layer, *i.e.,* replacing Eqn. (2) by $\widetilde{\boldsymbol{x}}_i = \boldsymbol{x}_i \cdot \boldsymbol{W}_1 - \boldsymbol{b}$, (2) $\underline{w/o \text{ MoE}}$ without MoE-enhanced adaptor ($G = 1$ in Eqn. (3)), (3) $\underline{w/o \text{ S-I}}$ without sequence-item contrastive task, and (4) $\underline{w/o \text{ S-S}}$ without sequence-sequence contrastive task.

The experimental results of the proposed approach UniSRec and its variants are reported in Figure 3. We can observe that all the proposed components are useful to improve the recommendation performance. The variant *w/o* MoE has poor performances because the MoE-enhance adaptor is the key component to improve the representation capacity for domain fusion and adaptation.

3.3.3 Performance Comparison w.r.t. Long-tail Items. One motivation to learn universal and transferable sequence representations is to alleviate the cold-start recommendation issue. To verify this, we split the test data into different groups according to the popularity of ground-truth items in the training data, and then compare the improved ratio of Recall@10 score (*w.r.t.* the baseline SASRec) in each group. From Figure 4, we can observe that the proposed approach outperforms the other baseline models in most cases, especially when the ground truth item is unpopular, *e.g.,* group [0, 5) on Industrial and Scientific datasets and group [0, 20) on Online Retail dataset. The results show that long-tail items can benefit from the learned universal sequence representations.

3.4 Case Study

As shown in Table 3, we can see that our approach can achieve good performance in a cross-platform setting (from *Amazon* to *Online Retail*). To our knowledge, in the literature of recommender systems, there are few studies that can conduct cross-platform recommendation [34]. It is interesting to study what kind of knowledge is actually transferred across different platforms. For this purpose, we present an illustrative case in Figure 5.

This example presents two short sequences of two different users from Amazon and Online Retail, respectively. It can be observed that our approach doesn’t rely on explicit item IDs to capture the user preference. Instead, it tries to capture the semantic associations by modeling the sequential patterns. Especially, the two short sequences correspond to a semantic transition from the keyword of “*dog*” to the keyword “*cat*” as shown in the item title. It shows that our approach can capture universal sequential patterns across platforms in terms of general textual semantics.

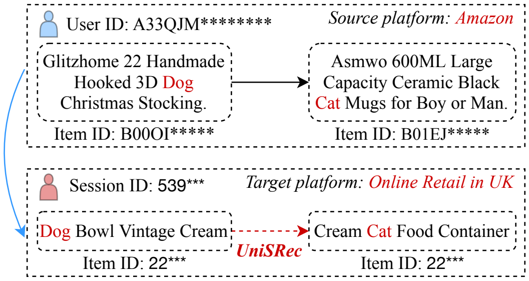

Figure 5: The purchase history of a user in the source platform (top) and the purchase history within an anonymous session in the target platform (bottom). There are naturally no overlapping users and items between the two platforms. This case shows that UniSRec can capture the universal semantic sequence pattern (*e.g.,* “Dog → Cat”) from large-scale pre-training, which helps improve the recommendation performance of the target platform.

Note that there might be also semantic correlations among other keywords. Here, we select the keywords of “dog” and “cat” just for simplifying the illustration. We leave a deep investigation of the universal representations as future work.

4 RELATED WORK

Sequential recommendation. To alleviate information overload, recommender systems are widely studied and deployed in various real-world services. Specially, it has been shown that sequential behaviors are important signals to reflect user preferences, and thus sequential recommendation has recieved much attention from both research and industry community [6, 18]. Early works on this topic adopt the Markov Chain assumption by estimating item-item transition probability matrices [18]. With the development of deep learning, Hidasi et al. [6] firstly introduce Gated Recurrent Units (GRU) to model sequential behaviors. Then various kinds of neural network techniques are proposed to encode interaction sequences, such as Recurrent Neural Network (RNN) [13], Transformer [5, 7, 10, 22], Multilayer Perceptron (MLP) [37] and Graph Neural Network (GNN) [2, 27]. Besides directly mining sequential patterns of IDs, some works are proposed to model sequences with rich features [32, 36] or additional self-supervised signals [29, 30, 36]. However, item representations and model parameters of these methods are usually restricted to specific data domains or platforms, making it difficult to leverage multi-domain data to improve sequential recommendation.

Transfer learning in recommender systems. To deal with the data sparsity and cold-start issues in the recommender systems, various works aim to leverage behavior information from other domains [39] or platforms [14] to improve recommendation performance of the target domain or both domains [38]. Most works rely on explicit overlapping data for conducting the transfer across domains, such as common users [8] or items [20, 38], social networks [14] and attributes [25]. Recently, some works attempt to learn universal user representations for different user-oriented downstream tasks [31]. However, the need of explicitly shared data still limits the application scope of the above methods. In this work, we propose to represent items in a universal way via text-based PLMs for domain adaptation. Different from the recently proposed zero-shot recommendation methods [4], we do not require that the source and target domains are closely related. With carefully designed universal pre-training tasks and MoE-enhanced adaptor architecture, the proposed universal SRL approach can be pre-trained on data from multiple source domains, and further generalize to different domains without explicitly shared anchors.

5 CONCLUSIONS

In this paper, we propose the universal sequence representation learning approach for recommender systems, named *UniSRec*. Different from existing sequential recommendation methods that reply on explicit item IDs for representation learning, the proposed approach UniSRec utilizes item texts to learn more transferable representations for sequential recommendation. Specifically, we design a lightweight architecture based on parametric whitening and MoE-enhanced adaptor to learn the universal item representations. We further design two contrastive pre-training tasks to learn universal sequence representations from multi-domain sequences. For future work, we will consider collecting more recommendation data to train larger user behavior models. Besides, we will explore more kinds of side information to represent items and improve sequence representation learning, *e.g.,* images and videos.

ACKNOWLEDGMENTS

This work was partially supported by National Natural Science Foundation of China under Grant No. 61872369, Beijing Natural Science Foundation under Grant No. 4222027, and Beijing Outstanding Young Scientist Program under Grant No. BJJWZYJH012019100020098. This work was partially supported by Beijing Academy of Artificial Intelligence (BAAI). Yupeng Hou was also partially supported by Alibaba Group through Alibaba Research Intern Program. Xin Zhao is the corresponding author.

REFERENCES

[1] Yoshua Bengio, Aaron C. Courville, and Pascal Vincent. 2013. Representation Learning: A Review and New Perspectives. *TPAMI* (2013).

[2] Jianxin Chang, Chen Gao, Yu Zheng, Yiqun Hui, Yanan Niu, Yang Song, Depeng Jin, and Yong Li. 2021. Sequential Recommendation with Graph Neural Networks. In *SIGIR*.

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL*.

[4] Hao Ding, Yifei Ma, Anoop Deoras, Yuyang Wang, and Hao Wang. 2021. Zero-Shot Recommender Systems. *arXiv preprint arXiv:2105.08318* (2021).

[5] Zhankui He, Handong Zhao, Zhe Lin, Zhaowen Wang, Ajinkya Kale, and Julian McAuley. 2021. Locally constrained self-attentive sequential recommendation. In *CIKM*.

[6] Balázs Hidasi, Alexandros Karatzoglou, Linas Baltrunas, and Domonkos Tikk. 2016. Session-based Recommendations with Recurrent Neural Networks. In *ICLR*.

[7] Yupeng Hou, Binbin Hu, Zhiqiang Zhang, and Wayne Xin Zhao. 2022. CORE: Simple and Effective Session-based Recommendation within Consistent Representation Space. In *SIGIR*.

[8] Guangneng Hu, Yu Zhang, and Qiang Yang. 2018. CoNet: Collaborative Cross Networks for Cross-Domain Recommendation. In *CIKM*.

[9] Junjie Huang, Duyu Tang, Wanjun Zhong, Shuai Lu, Linjun Shou, Ming Gong, Daxin Jiang, and Nan Duan. 2021. WhiteningBERT: An Easy Unsupervised Sentence Embedding Approach. In *Findings of EMNLP*.

[10] Wang-Cheng Kang and Julian J. McAuley. 2018. Self-Attentive Sequential Recommendation. In *ICDM*.

[11] Bohan Li, Hao Zhou, Junxian He, Mingxuan Wang, Yiming Yang, and Lei Li. 2020. On the Sentence Embeddings from Pre-trained Language Models. In *EMNLP*.

[12] Chenglin Li, Mingjun Zhao, Huanming Zhang, Chenyun Yu, Lei Cheng, Guoqiang Shu, Beibei Kong, and Di Niu. 2022. RecGURU: Adversarial Learning of Generalized User Representations for Cross-Domain Recommendation. In *WSDM*.

[13] Jing Li, Pengjie Ren, Zhumin Chen, Zhaochun Ren, Tao Lian, and Jun Ma. 2017. Neural Attentive Session-based Recommendation. In *CIKM*.

[14] Tzu-Heng Lin, Chen Gao, and Yong Li. 2019. CROSS: Cross-platform Recommendation for Social E-Commerce. In *SIGIR*.

[15] Zihan Lin, Changxin Tian, Yupeng Hou, and Wayne Xin Zhao. 2022. Improving Graph Collaborative Filtering with Neighborhood-enriched Contrastive Learning. In *TheWebConf*.

[16] Shanlei Mu, Yupeng Hou, Wayne Xin Zhao, Yaliang Li, and Bolin Ding. 2022. ID-Agnostic User Behavior Pre-training for Sequential Recommendation. *arXiv preprint arXiv:2206.02323* (2022).

[17] Jianmo Ni, Jiacheng Li, and Julian J. McAuley. 2019. Justifying Recommendations using Distantly-Labeled Reviews and Fine-Grained Aspects. In *EMNLP*.

[18] Steffen Rendle, Christoph Freudenthaler, and Lars Schmidt-Thieme. 2010. Factorizing personalized Markov chains for next-basket recommendation. In *WWW*.

[19] Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc V. Le, Geoffrey E. Hinton, and Jeff Dean. 2017. Outrageously Large Neural Networks: The Sparsely-Gated Mixture-of-Experts Layer. In *ICLR*.

[20] Ajit Paul Singh and Geoffrey J. Gordon. 2008. Relational learning via collective matrix factorization. In *SIGKDD*.

[21] Jianlin Su, Jiarun Cao, Weijie Liu, and Yangyiwen Ou. 2021. Whitening Sentence Representations for Better Semantics and Faster Retrieval. *arXiv preprint arXiv:2103.15316* (2021).

[22] 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 *CIKM*.

[23] Hongyan Tang, Junning Liu, Ming Zhao, and Xudong Gong. 2020. Progressive Layered Extraction (PLE): A Novel Multi-Task Learning (MTL) Model for Personalized Recommendations. In *RecSys*.

[24] Jiaxi Tang and Ke Wang. 2018. Personalized Top-N Sequential Recommendation via Convolutional Sequence Embedding. In *WSDM*.

[25] Jie Tang, Sen Wu, Jimeng Sun, and Hang Su. 2012. Cross-domain collaboration recommendation. In *SIGKDD*.

[26] 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 *NeurIPS*.

[27] Shu Wu, Yuyuan Tang, Yanqiao Zhu, Liang Wang, Xing Xie, and Tieniu Tan. 2019. Session-Based Recommendation with Graph Neural Networks. In *AAAI*.

[28] Ruobing Xie, Qi Liu, Liangdong Wang, Shukai Liu, Bo Zhang, and Leyu Lin. 2022. Contrastive Cross-domain Recommendation in Matching. In *SIGKDD*.

[29] Xu Xie, Fei Sun, Zhaoyang Liu, Shiwen Wu, Jinyang Gao, Bolin Ding, and Bin Cui. 2022. Contrastive learning for sequential recommendation. In *ICDE*.

[30] Junliang Yu, Hongzhi Yin, Xin Xia, Tong Chen, Jundong Li, and Zi Huang. 2022. Self-Supervised Learning for Recommender Systems: A Survey. *arXiv preprint arXiv:2203.15876* (2022).

[31] Fajie Yuan, Xiangnan He, Alexandros Karatzoglou, and Liguang Zhang. 2020. Parameter-Efficient Transfer from Sequential Behaviors for User Modeling and Recommendation. In *SIGIR*.

[32] 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 *IJCAI*.

[33] Wayne Xin Zhao, Junhua Chen, Pengfei Wang, Qi Gu, and Ji-Rong Wen. 2020. Revisiting Alternative Experimental Settings for Evaluating Top-N Item Recommendation Algorithms. In *CIKM*.

[34] Wayne Xin Zhao, Yanwei Guo, Yulan He, Han Jiang, Yuexin Wu, and Xiaoming Li. 2014. We know what you want to buy: a demographic-based system for product recommendation on microblogs. In *KDD*.

[35] Wayne Xin Zhao, Shanlei Mu, Yupeng Hou, Zihan Lin, Yushuo Chen, Xingyu Pan, Kaiyuan Li, Yujie Lu, Hui Wang, Changxin Tian, Yingqian Min, Zhichao Feng, Xinyan Fan, Xu Chen, Pengfei Wang, Wendi Ji, Yaliang Li, Xiaoling Wang, and Ji-Rong Wen. 2021. RecBole: Towards a Unified, Comprehensive and Efficient Framework for Recommendation Algorithms. In *CIKM*.

[36] 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*.

[37] Kun Zhou, Hui Yu, Wayne Xin Zhao, and Ji-Rong Wen. 2022. Filter-enhanced MLP is All You Need for Sequential Recommendation. In *TheWebConf*.

[38] Feng Zhu, Chaochao Chen, Yan Wang, Guanfeng Liu, and Xiaolin Zheng. 2019. DTCDR: A Framework for Dual-Target Cross-Domain Recommendation. In *CIKM*.

[39] Feng Zhu, Yan Wang, Chaochao Chen, Jun Zhou, Longfei Li, and Guanfeng Liu. 2021. Cross-Domain Recommendation: Challenges, Progress, and Prospects. In *IJCAI*.