Variational Session-based Recommendation Using Normalizing Flows

Fan Zhou
University of Electronic Science and Technology of China, Chengdu, China
fan.zhou@uestc.edu.cn

Zijing Wen
University of Electronic Science and Technology of China, Chengdu, China
wenzijing93@gmail.com

Kunpeng Zhang
University of Maryland, College park
kpzhang@umd.edu

Goce Trajcevski
Iowa State University, Ames
gocet25@iastate.edu

Ting Zhong∗
University of Electronic Science and Technology of China, Chengdu, China
zhongting@uestc.edu.cn

ABSTRACT
We present a novel generative Session-Based Recommendation (SBR) framework, called Variational Session-based Recommendation (VASER) – a non-linear probabilistic methodology allowing Bayesian inference for flexible parameter estimation of sequential recommendations. Instead of directly applying extended Variational AutoEncoders (VAE) to SBR, the proposed method introduces normalizing flows to estimate the probabilistic posterior, which is more effective than the agnostic presumed prior approximation used in existing deep generative recommendation approaches. VASER explores soft attention mechanism to upweight the important clicks in a session. We empirically demonstrate that the proposed model significantly outperforms several state-of-the-art baselines, including the recently-proposed RNN/VAE-based approaches on real-world datasets.

CCS CONCEPTS
• Information systems → Recommender systems; • Computing methodologies → Neural networks.

KEYWORDS
Session-based recommendation, variational autoencoders, normalizing flows

ACM Reference Format:

1 INTRODUCTION
Session-based recommendation (SBR) [18, 36] aims at predicting user’s next action based on a series of recent actions. It is a kind of sequence learning/recommendation task where longer-term user historical activities are usually unavailable and the recommendations need to be made in accordance with an assumed short-term interests of the (anonymous) user. It has been widely used in various applications, such as news recommendation, e-commerce, video and classified advertisement recommendation [10, 36].

Recent advances in deep learning have spurred the use of recurrent neural networks (RNNs) based methods to model SBR [17, 18, 28, 33], achieving significant improvement on recommendation accuracy over traditional sequence-based models such as factorizing personalized Markov chains (FPMC) [16, 38] and feature-based matrix factorization (MF) [5, 51]. Specifically, GRU4Rec [18] – a first application of augmented gated recurrent units (GRUs) [8] – was developed to address SBR by encoding user’s preference and learning it for next-click prediction. Subsequently, a few improvements to GRU4Rec have been proposed – e.g., incorporating attention mechanism [28]; employing hierarchical recurrent networks [28]; augmenting data with additional features associated with items [19]; prioritizing short attention/memory [33]; and introducing more sophisticated ranking algorithms [17].

Complementary to these works, recent efforts on incorporating stochastic latent variables trained by deep generative models (e.g., variational autoencoders (VAE) [25, 40]) have enabled significant progress in many natural language processing tasks (e.g., dialogue generation and machine translation [2, 3, 11, 14, 21]). Similarly, various generative models including VAEs have demonstrated potential for learning effective non-linear representations of user-item interactions [7, 23, 27, 30, 32] in collaborative filtering settings. For the most part, they either model the generation process of auxiliary information (e.g., content and ratings) [7, 27, 30] or build a probabilistic latent-variable framework that shares statistical strength among users and items [7, 23, 32].

Despite the improvements over conventional item recommendation, the aforementioned Bayesian models cannot be directly generalized to SBR due to the following reasons:
(1) Data availability: the lack of users’ profile information and long-term interaction data makes these models not work well in SBR settings.
(2) Bypassing issue: autoregressive models (e.g., LSTM [20] and GRU [8]) combined with the soft attention mechanisms [1] have capabilities of reconstructing an encoded session on their own. This
may weaken the effects of the incorporated latent factors [3], which can potentially reduce the performance of the VAE-based models. 

3 Bias inference: VAE based models usually assume a predefined prior for latent factors [24], e.g., multivariate Gaussian which might result in the inferred approximate posterior greatly deviating from the true distribution.

We extend VAEs to model implicit feedbacks of user-item interactions in a session, and present the Variational Session-based Recommendation (VASER). While retaining the Bayesian inference of VAEs and enabling exploration of non-linear probabilistic latent-variable models, the VASER model: (1) effectively addresses the problem of unimodal and simple parametric problems of existing SBR methods; and (2) largely ameliorates the bias inference problem of existing VAE based recommendation methods. Specifically, we make the following contributions:

• VASER augments the RNNs based SBR models with stochastic latent variables, enabling stable and effective approximate inference of a high-level “objective” of an entire session from the observed clicks.

• We exploit the flows to approximate the real posterior of stochastic latent factors, which can largely alleviate the inference bias in existing VAE based recommendation models and improve the next click prediction accuracy.

• We demonstrate that VASER achieves improvements in SBR performance over the baselines on several real-world datasets.

2 PRELIMINARIES

We now introduce the basic notation used in the rest of the paper and formally define the problem. Recalling that SBR aims at predicting which item an anonymous user would like to click next given his current sequence of interaction data, we also provide an overview of the RNN based approaches.

2.1 Problem Definition

Formally, we have a set of sessions $S$, and each session $s_i \in S$, consists of a sequence of user actions (e.g., click, purchase, etc.): $s_i = [x_{i1}, \cdots, x_{iN}]$ (interchangeably denoted by $x_{i1:N}$), where $x_{ij} \in \mathbb{R}$ ($1 \leq j \leq N$, $N$ is the length of the session). $s_i$ is an interaction with item $j$ in the session, assumed to be mapped to the domain $\mathbb{R}$. When no ambiguity arises, we will omit the index of the session – i.e., they train item embedding along with calculating loss of training sessions. Therefore, this type of supervised training may indeed “memorize” the sequential information of a given session, which may be “conducted” in the testing phase as the items in testing sessions would look-up the embedding matrix. As observed in the experiments in [30], this dynamic embedding method may significantly improve the performance.

An important observation is that all these works train the model in an explicit autoregressive fashion, i.e., they split the sessions (both training and testing) into a set of sub-sessions. Thus, a session $s = [x_1, \cdots, x_N]$ would be divided into $N - 1$ sub-sessions: $s^1 = [x_1, \cdots, x_{N-1}], s^2 = [x_1, \cdots, x_{N-2}], \cdots, s^{N-1} = [x_1, x_2]$ and the original session $s$ - all of which would be fed into the models for training or testing. Although not explicitly specified, this kind of autoregressive training improves the overall performance of the models, since a longer session actually contains (and thus “memorizes”) the sub-sessions. We note that this autoregressive training trick has also been explored in recent CNN based SBR models [45, 52].

3 MAIN METHODOLOGIES

We propose the VASER model, as illustrated in Figure 1, consisting of two main components, namely GRU module and attention module. The GRU module captures sequential preferences, and the hidden state can exploit the non-linear preferences. The attention module is used to enhance the GRU network by dynamically selecting and linearly combining different parts of the input sequence. VASER employs a deterministic attention mechanism. VASER incorporates the normalizing flows for flexible posterior approximation. In the sequel, we present the general framework of VASER with theoretical background and training procedure.

3.1 Session Generative Model

We consider a click session generative process as follows. For each session $s = [x_1, \cdots, x_N]$, the model samples $d$-dimensional latent representation from an appropriate prior distribution $p(z)$. The latent factor $z$ is then transformed via a non-linear function $f_{\theta}(z)$ – a suitable likelihood function parameterized by $\theta$ – to produce a probability distribution $\pi(z)$ (e.g., a multinomial distribution) over
where each \( z_k \in f \) is invertible (i.e., \( z_{k-1} = f^{-1}_k(z_k) \)), and according to the definition of probability \( \int p(z_k)dz_k = \int p_{k-1}(z_{k-1})dz_{k-1} = 1 \), for a collection of variables \( z = \{ z_0, \ldots, z_K \} \), we can obtain the distributions \( p_K(z_K) \) more flexibly:

\[
p_K(z_K) = p(z_0) \det \frac{dz_K}{dz_0}^{-1}
\]

where \( \det \frac{dz}{dx} \) is the Jacobian determinant of \( f \).

The path traversed by the random variables \( z_k = f_k(z_{k-1}) \) with initial distribution \( p(z_0) \) is called the flow, and the whole path formed by the successive distributions \( p_K(z_K) \) refers to a normalizing flow. To ensure Eq.(6) is tractable, it should satisfy that (a) the transformation \( f_k \) must be easy to invert, and (2) the determinant of its Jacobian is easy to compute [39]. The two constraints allow the transformation to be made deeper by composing multiple instances of it, and the result will still be a valid normalizing flow. Now the log-likelihood of approximate posterior \( q_K(z_K|s) \) can be computed iteratively by using the log on both sides of Eq.(6)

\[
\log q_K(z_K|s) = \log q_0(z_0|s) - \sum_{k=1}^K \log \det \left| \frac{dz_k}{dz_{k-1}} \right|
\]

where the base distribution \( z_0 \sim q_0(z_0|s) \) is a Gaussian in our implementation.

One of the flow transformations is the planar flow introduced in [39], given by:

\[
f(z) = z + u \sigma(w^Tz + b)
\]

where \( u, w \in \mathbb{R}^d \) and \( b \in \mathbb{R} \) are parameters, and \( \sigma \) is a suitable smooth non-linear activation function (e.g., tanh). According to the Matrix determinant lemma, the Jacobian of this transformation is:

\[
\det \frac{df}{dz} = \left| 1 + u^T \sigma'(w^Tz + b)w \right|
\]

where \( \sigma' \) is the derivative activation and can be computed in \( O(d) \) time — \( d \) is the dimension of \( z \).

In this paper, we use the planar flow as the invertible transformation for its simplicity and efficiency. We parameterize the approximation posterior \( q_K(z_K|s) \) with a flow, i.e.,

\[
q_K(z_K|s) = q(z_K)
\]

the ELBO of Eq.(2) can be modified as

\[
\mathcal{L}(s; \theta, \phi) = \mathbb{E}_{q_0(s)} \left[ \log p_0(z_k|s) \right] - \mathbb{E}_{q_0(s)} \left[ \log q_0(s) \right] + \beta \mathbb{E}_{q(s_0)} \left[ \log p_0(z_K|s) \right]
\]

\[
+ \beta \mathbb{E}_{q(s_0)} \left[ \sum_{k=1}^K \log \det \left| \frac{dz_k}{dz_{k-1}} \right|^{-1} \right]
\]
where the first term is trained to reconstruct the sessions; the second term is a constant; and the last two terms are the flows. The coefficient $\beta$ is a regularizer of the flows, which is very similar to the annealing factor for regularizing KL-divergence [3].

4 EXPERIMENTAL OBSERVATIONS

We now describe the experimental settings and report the empirical evaluation results.

Table 1: Statistics of the datasets.

<table>
<thead>
<tr>
<th>Datasets</th>
<th>YOOCHOOSE 1/64</th>
<th>YOOCHOOSE 1/4</th>
</tr>
</thead>
<tbody>
<tr>
<td>#clicks</td>
<td>557,248</td>
<td>832,6407</td>
</tr>
<tr>
<td>#train sessions</td>
<td>355,385</td>
<td>621,6184</td>
</tr>
<tr>
<td>#test sessions</td>
<td>52,956</td>
<td>56,616</td>
</tr>
<tr>
<td>#items</td>
<td>17,626</td>
<td>30,903</td>
</tr>
<tr>
<td>avg. session length</td>
<td>6.27</td>
<td>5.83</td>
</tr>
</tbody>
</table>

4.1 Datasets

For fair comparison, we evaluate different methods on a real-world transaction datasets YOOCHOOSE\textsuperscript{1}, which has been widely used for evaluating SBR approaches. Following previous works [28, 33, 44], we preprocess the primary data as follows: (1) We filter out sessions of length 1 and items that appear less than 5 times for the two datasets; (2) We respectively use the sessions of subsequent day for testing, and then filter out clicks from the test set where the clicked items did not appear in the training set; and (3) We sort the training sequences by time and train all models on more recent fractions (i.e., 1/64 and 1/4) of training sessions. Table 1 shows the statistics of the datasets.

4.2 Baselines

To demonstrate the effectiveness of our model, we conduct extensive comparisons to the following state-of-the-art methods:

- **Item-KNN** [42]: It is an item-to-item model that recommends items that are similar to previously visited items based on cosine similarity.
- **GRU4Rec\textsuperscript{2}** [18]: It is an RNN-based deep learning model for session-based recommendation. It employs GRU units to capture sequential patterns and utilizes session-parallel minibatching trick and ranking-based loss functions during the training.
- **NARM\textsuperscript{3}** [28]: It is an RNN-based model employing (deterministic) attention mechanism to capture main purpose from the hidden states and combines it with the sequential behavior as the final representation to generate recommendations.
- **STAMP** [33]: It is a priority model which captures users’ general interests from the long-term memory of a session context, and current interests from the short-term memory of recent clicks.
- **ReLaVaR** [4]: It is a Bayesian version of GRU4Rec which treats the network recurrent units as stochastic latent variables with some prior distributions and infers the corresponding posteriors for prediction and recommendation generation. This is an item-level variational inference based SBR method which uses independent Gaussian as the prior for items.
- **VRM** [49]: It is a recent proposed method directly applying VAE on session-based recommendation. Unlike ReLaVaR, an item-level variational method, VRM models the stochastic inference on the session-level.

4.3 Metrics

Following previous works [16, 17, 28, 33], the primary evaluation metric is Recall@20 – i.e., the proportion of cases having the desired item falling into the top-20 predicted items in all test cases. Note that the Recall score is equal to the Hit-Precision score used in [33]. The second metric is MRR@20 (Mean Reciprocal Rank) – i.e., the average of reciprocal ranks of the desired items. The reciprocal rank is set to zero if the rank is lower than top-20. MRR takes into account the rank of the item, which is important when the order of recommendations matters. Note that the higher the Recall@20 and MRR@20, the better the performance.

4.4 Settings

For all methods, the embedding size of items is set to 50. The number of hidden units in GRU layer is set to 100. All models are trained with Adam and the mini-batch size is fixed at 512. Following [28, 33], we truncated BPTT using a fixed window of 30 time-steps for the two YOOCHOOSE datasets. Also following [28, 33], 10% of the training data are used as the validation set. For the VASER model, parameters $d, K$ and $\beta$ are respectively 100, 16 and 0.2, if not specified.

4.5 Overall Performance

Table 2: Performance comparison among all session-based recommendation methods over two datasets.

<table>
<thead>
<tr>
<th></th>
<th>YOOCHOOSE 1/64</th>
<th>YOOCHOOSE 1/4</th>
</tr>
</thead>
<tbody>
<tr>
<td>item-RNN</td>
<td>53.12</td>
<td>52.43</td>
</tr>
<tr>
<td>GRU4REC</td>
<td>62.40</td>
<td>59.58</td>
</tr>
<tr>
<td>NARM</td>
<td>70.13</td>
<td>69.75</td>
</tr>
<tr>
<td>STAMP</td>
<td>70.21</td>
<td>70.45</td>
</tr>
<tr>
<td>ReLaVaR</td>
<td>64.32</td>
<td>60.53</td>
</tr>
<tr>
<td>CRM</td>
<td>69.32</td>
<td>68.22</td>
</tr>
<tr>
<td>VASER</td>
<td>71.85</td>
<td>70.74</td>
</tr>
</tbody>
</table>

4.5.1 Comparison against SBR baselines (Q1). Table 2 shows the results of comparison to the existing state-of-the-art SBR methods, from which we can clearly observe that the proposed model perform the best on two metrics throughout two datasets.

Overall, the RNN based methods, including ours, consistently outperform the traditional baselines, which demonstrates that autoregressive models are good at learning sequential user click behaviors.
Nevertheless, RNN models alone cannot deal with complicate user-click sessions which usually have unintended clicks and/or contain one or more browse themes. This problem can be largely overcome by incorporating the attention mechanism in recent methods like NARM and STAMP. The most recent work GRU4Rec++ does not exhibit expected results on the two datasets, regardless that it can improve their original method (GRU4Rec) with the sampling trick. This result also proves one of our motivations that autoregressive models are constrained with their capability of modeling sparse and high-dimensional data.

By modeling session generation in a probabilistic generative latent variable framework, our model outperform the best baseline (either NARM or STAMP) by a significant margin. Note that in our reimplementation, the two baselines (NARM or STAMP) exhibit higher scores than their original reporting on two YOOCHOOSE datasets.

The benefit of VASER can be visualized in Figure 2, where the encoding space of NARM and VASER is plotted with t-SNE. Recall that both methods predict the last item $x_N$ based on the learned representation of prefix session $s' = \{x_1, \cdots, x_{N-1}\}$. The main difference is that the encoding space of NARM is the concatenation of deterministic attention $c$ and the last hidden state $h_{N-1}$ of GRU, while the encoding space of VASER is the combination of the posterior of hidden state $q(z_k)$ and the posterior of attention $q(c_k)$. Apparently, VASER explores more space for encoding the sessions and exhibits more scattered distribution. The benefits of such encoding can be understood intuitively, i.e., the more inseparable the sessions, the more difficult for the models to discriminate the spatial adjacent ones, which, consequently, are more prone to making wrong predictions.

We also investigated the impact of session length on the recommendation performance. Intuitively, the longer the sessions, the worse the prediction performance on average. The results on YOOCHOOSE 1/64 are shown in Figure 3 (results on YOOCHOOSE 1/4 are consistent, but omitted due to the lack of space), whereby we compared to the two best baselines. Our model slightly improve the recommendation performance over the baselines. However, we argue that due to the vanishing problem of autoregressive model, it is hard for RNNs-based methods to further improve the performance on modeling extremely long-term dependencies.

4.5.2 Effect of Components in VASER (Q2). By comparing to the methods directly applying VAE on SBR, both item-level and session-level, we can clearly see the importance of the flow based posterior approximation used in VASER.

Another important observation is that directly applying VAE on modeling items or sessions is not competitive. ReLaVaR, operating stochastic inference on item-level, is less effective than VRM, which models sessions in a variational seq2seq manner. Note that we omit comparison with CVRM, a variant of VRM, which takes the category information into account, due to the extremely sparse category labels on the YOOCHOOSE dataset. In fact, the category information plays a less important role in improving the recommendation performance according to the results in [49]. Although allowing Bayesian inference, the two models may incur larger inference gaps and underfitting problem due to the amortized inference alone used for posterior approximation [9]. This is in accordance with the findings in modeling language with vanilla VAEs [3, 21], i.e., the autoregressive models are powerful enough to decode the entire sequence, resulting in uselessness of stochastic latent factors. More importantly, these methods approximate an improper assumed distribution $q_\phi(z/s)$, e.g., the choice of diagonal-covariance Gaussian in [3, 21], and thus are subjected to heavy bias inference problem, as explained in Sec. 3.2. In contrast, our VASER model can largely alleviate this problem benefitting from the normalizing flows with flexible posterior approximation.

4.6 Impact of Parameters (Q3)

There are two important factors affecting the performance of the VASER model, i.e., the coefficient $\beta$ regularizing the flows and the determinant of Jacobian matrix, and $K$, the number of invertible transformations.

Figure 4 shows the impact of $\beta$ on VASER, where $\beta$ is gradually annealed to the value of 0.1, 0.2, 0.5 and 1. We observe in our experiments that the flow terms are usually ordered larger than the reconstruction term. Without annealing or annealing to a larger value, the performance of VASER model are not appealing, and even experience overfitting problem. On the contrary, if the value of $\beta$ is too small (e.g., below 0.2, the flows does not take effect as the decoder RNN will make the model converge, when the model rely less on the latent factors. As a consequence, there is a significant performance decline. As we explained earlier, this is caused by the overpower performance of RNN decoder. In addition to cost annealing, another possible way of alleviating this problem is to replace RNN with the dilated CNN suggested by [21].

Figure 5 investigates the impact of $K$ on two datasets. The planar flows used in VASER modifies the initial density by applying a series of deterministic attention transformations.
Learning approaches can also be adapted to solve the session-based recommendation problem, which incurs MC based models. Several works have been proposed to improve GRU4Rec with various models. For example, NARM and EDRec [28, 34] employ soft attention mechanism [1] to capture the user’s main purpose in the current session, which is combined with the last hidden state of GRU to compute the recommendation scores for each candidate item. Hidasi et al. [17] improved their GRU4Rec model by introducing tailored ranking loss functions.

5.2 Deep Generative Recommendation

Although there exist many deep recommendation models as mentioned above, relatively few works in the literature focus on applying generative models in the recommendation systems. Previous autoencoder based models [29, 41, 43, 46, 47, 50] show promising performance but are restricted to learning representation of items, and thus are difficult for Bayesian inference due to lack of Bayesian nature or high computational cost. Several recent works extend the ideas of applying VAEs to CF-based recommendation but mainly focus on combining various auxiliary features [7, 23]. The most related work are ReLaVaR [4] and VRM/CVRM [49], both of which apply VAE on the SBR tasks. ReLaVaR is an item-level stochastic inference method while VRM/CVRM are modeling session in a stochastic seq2seq manner. As sequential VAE models, they can be considered as directly applying VAEs in the SBR scenario.

Compared to existing sequential recommendation approaches, we model the problem within a probabilistic recommendation setting which allows our model for Bayesian inference. In addition, we derive novel flow-based model tailored for SBR task with the flexible posterior approximation, rather than presumed Gaussian distribution in previous work.

6 CONCLUSIONS AND FUTURE WORK

We presented VASER, a flow based generative framework for learning sequential click patterns. A distinct feature of the proposed model implementing VASER is that it enables learning non-linear interactions between user-clicks while allowing Bayesian inference. As demonstrated by the experiments, VASER achieves significant improvements for the session-based recommendation problem in comparison to existing methods. One of the most important implications of the results from this work is that instead of using amortized inference as in existing collaborative/sequential variational recommendation methods, flow based techniques could effectively improve the density approximation and deserves more attention in the recommendation community. In our future work, we are planning to focus on augmenting VASER to consider auxiliary information – e.g., coupling sequential information with other related contexts (category, price and click time), and on tackling the overall efficiency (e.g., by incorporating CNNs).

ACKNOWLEDGEMENTS

This work was supported by National Natural Science Foundation of China (Grants No.61602097 and No.61472064), NSF Grants III 1213038 and CNS 1646107, and an ONR grant N00014-14-10215.
REFERENCES


mendations. In ICDM.


[26] Rahul G Krishnan, Dawen Liang, and Matthew D Hoffman. 2018. On the chal-
lenges of learning with inference networks on sparse, high-dimensional data. In AISTATS.


[29] Sheng Li, Jaya Kawale, and Yun Fu. 2015. Deep Collaborative Filtering via Marginalized Denoising Auto-encoder. In CIKM.


[35] Fei Mi and Boi Faltings. 2018. Context Tree for Adaptive Session-based Recom-


[38] Steffen Rendle, Christoph Freudenthaler, and Lars Schmidt-Thieme. 2010. Factor-
izing personalized Markov chains for next-basket recommendation. In WWW.


[44] Yong Kiam Tan, Xinxing Xu, and Yong Liu. 2016. Improved Recurrent Neural Networks for Session-based Recommendations. In DLRS@RecSys.


[48] Pengfei Wang, Jiafeng Guo, Yanyan Lan, Jun Xu, Shengxian Wan, and Xueqi Cheng. 2015. Learning Hierarchical Representation Model for NextBasket Recom-
mendation. In SKAIR.


[50] Yao Wu, Christopher DuBois, Alice X Zheng, and Martin Ester. 2016. Collabora-
tive Denoising Auto-Encoders for Top-N Recommender Systems. In WSDM.


[52] Fajie Yuan, Alexandros Karatzoglou, Ioannis Arapakis, Joemon M Jose, and Xing-an He. 2018. A Simple but Hard-to-Beat Baseline for Session-based Recom-