Noname manuscript No.
(will be inserted by the editor)

Dual Self-supervised Variational Autoencoder for Collaborative Filtering

Jing Wang, Jun Wu(✉)
Caiyan Jia, Zhifei Zhang

Received: date / Accepted: date

Abstract Variational autoencoder (VAE) is considered as an emerging model for ensuring competitive performance in recommender systems. However, existing VAE models may fail to provide satisfactory recommendation results in presence of highly sparse user-item interactions. In this paper, we propose a Dual Self-supervised Variational Autoencoder (DSVAE) model to improve both the generalization ability of VAE model on the sparse interaction datasets and personalized characteristic of recommendation results. Specifically, we supplement the classical supervised task of recommendation with dual self-supervised tasks which contains two SSL tasks. One is to align the representations learned from different views, where views are generated by data augmentation, and the other is to discriminate reconstructed feedback data from the rest of user input feedback data. Furthermore, we aim to optimize a combined objective of recommendation task and pretext task, making them to reinforce each other during the learning process. Extensive experiment results on three real-world benchmarks validate the superiority of our DSVAE model to state-of-the-art VAE style recommendation techniques.

Keywords Recommender Systems · Self-supervised Learning · Variational Autoencoder

1 Introduction

Nowadays, recommender systems have been an indispensable tool in various online applications, including E-commerce platforms, video recommendation, for different users in the situations of information overload. The core recommendation method in RSs is collaborative filtering (CF) [1], which analyzes the

Jing Wang, Jun Wu, Caiyan Jia, Zhifei Zhang
School of Computer and Information Technology, Beijing Jiaotong University, Beijing, China
E-mail: {19125244, cyjia, zhfzhang, wuj}@bjtu.edu.cn

users and interdependencies among items, with the goal of identifying unobserved user-item associations. As dominant models in CF, latent factor model and matrix factorization model [3, 8, 10, 21, 20] gain substantial attention due to their simplicity and effectiveness. However, its performance is limited by inferior expressiveness of interaction data.

One of promising solutions to improve prediction is applying non-linear neural-based approaches to enhance expressiveness of models [7, 26], and an emerging method is Variational Autoencoder (VAE) [9, 12, 26]. It is a powerful generative model which generates new data by modeling the underlying probability distribution of input data so that the diversity of recommendations can be controlled by sampling multiple results from that distribution. When the observed user-item interactions are much less than the unobserved ones, VAE based CF model faces the challenge of data sparsity which leads to unsatisfied recommendation results. Incorporating content information to augment VAE is a promising solution to the problem of data sparsity. Recently there exists a few tentative studies [4, 11, 18, 19, 27, 29] which encompass VAE through augmenting structures to model both content information and user interaction information. Content information includes textual reviews, social network, geotag and knowledge graphs. However, each kind of side information is with its own characteristics so that it is hard to devise a generic approach to exploit different kinds of side information. When VAE is used to generate recommendation results, we expect that user-item interaction data can offer auxiliary signal to mitigate data sparsity problem. Besides, we hope that the recommendation results generated for different users can reflect obvious differences in order to construct personalized recommendation.

In order to overcome the above shortcomings, in this paper, we propose Dual Self-supervised Variational Autoencoder (DSVAE), which mitigates data sparsity problem and encourages personalized characteristic of recommendation list for user. DSVAE seamlessly integrates dual self-supervised pretext task with main recommendation task by jointly optimizing a unified objective function, where dual self-supervised learning strategy can be applied to most of VAE based collaborative filtering models. Dual SSL strategy consists of two key self-supervised tasks: (1) with contrastive learning, exploring correlation between two views of the same user interaction data which are generated by data augmentation, and (2) introducing contrastive loss to distinguish each user reconstructed feedback vector from the rest of user input feedback vector. Dual self-supervised learning strategy can explore additional supervision signal in the context of user-item interaction sparsity and improve personalized characteristic of recommendation result. To the best of our knowledge, our DSVAE is the first model combining self-supervised learning with VAE based recommendation methods. Results from extensive experiments on three benchmarks show that DSVAE scheme outperforms several state-of-the-art interaction-only VAE based recommendation methods. Owing to dual self-supervised strategy, DSVAE achieves comparable performance with content aware VAE models.

The rest of this paper is organized as follows. Section 2 reviews related work. Section 3 elaborates the proposed DSVAE approach. Extensive experiments are conducted in Section 4, followed by the conclusion in Section 5.

2 Related Work

As a new collaborative filtering algorithm, DSVE can be seen as an integration of VAE and self-supervised learning (SSL).

2.1 Variational Autoencoder

Variational autoencoder [9, 12, 26] predicts recommendation item list for user by reconstructing input interaction data, presenting unique advantage in collaborative filtering in recent years. It is an approach that consists of two parts: inference model and generative model. The inference model encodes interaction data into feature representation by learning its distribution, then the generative model decodes feature representation to generate meaningful outputs. VAE uses a variational bayesian method for training with an optimisation objective containing the sum of reconstruction loss of input data and KL-divergence between variational posterior and prior distribution. Unobserved interactions can be obtained by looking at reconstructed user feedback vector.

One explanation for the good performance achieved by VAE on the collaborative filtering is its probabilistic nature [23]. VAE does not devote to extract deterministic feature representations, but rather learns distributions over these representations, allowing it to account for the uncertainty of latent space. Another advantage of VAE is that as all users share the same encoder/decoder network, the number of parameters required for VAE is independent of the number of users. [15] This is in contrast to some traditional latent factor collaborative filtering models, where a unique latent vector is learned for each user.

Roughly speaking, existing methods of VAE based recommendation methods can be divided into two categories. The first manner is concern about the architecture of VAE, including introducing proper prior, optimizing network of encoder or decoder [13, 14, 15, 16, 22, 23]. eg., Rec-VAE [22] achieved good performance and the main contribution of it is the introduction of composite prior distribution which is a mixture of standard Gaussian prior and the distribution of latent code from previous model iteration. With integrating the idea of clustering, MacridVAE [16] designed unique encoder and decoder network which learned disentangled representations to enhance robustness and interpretability of recommendation. Although effective, they still suffer from problem that they are are challenging to train on sparse user-item interactions. Thus, in order to mitigate the sparsity of user-item interactions, the other methods concentrates on augmenting VAE with blending various content information [4, 11, 18, 19, 27, 29]. eg., Split-Merge CVAE [18] expanded variational autoencoder with multiple condition labels, enforcing model to distinguish and cluster users in latent subspace. Although effective, side information is only accessible in some specific recommendation scenarios.

Unlike above methods, our approach focuses on exploring self-supervised learning to extract additional supervision signals to alleviate data sparsity problem as well as improve personalized characteristic of recommendation results.

2.2 Self-supervised Learning

Self-supervised learning generally contains two components: auxiliary task and loss function. The auxiliary task means that task being solved is not of genuine interest, but is solved only for the true purpose [6]. Loss function is to measure the difference between prediction and target which in auxiliary task. Various self-supervised learning tasks have been used in CV. For example, S3VAE [31] shuffled temporal order of sequential data, and trained a model to learn time-invariant representation to exclude any dynamic information. In NLP, masked language task is introduced in the BERT [2] model, to capture the dependencies among tokens.

Inspired by success of self-supervised learning, there have been many works introduced SSL to recommendation. SGL [25] generated multiple views of nodes on user-item graph and designed tasks to learn discriminative node representation. Another model [28] conducted self-supervised learning in training two-tower Deep Neural Net model for recommendations. It proposed two different data augmentation methods and leveraged contrastive learning to make sure the same examples after different augmentation could still be recognized exactly.

These methods research SSL on graph-based models and two-tower DNN models, achieving good performance. In this paper, DSVAE combines SSL with VAE based CF models. Remarkably, the strategy we proposed is specifically for most VAE based recommendation.

3 The Proposed Method

3.1 Notations

In this paper, we denote scalars, vectors, and matrices using lower case, bold lower case, and bold upper case letters, e.g., $n, \mathbf{x}, \mathbf{X}$. Let $\mathbf{X} \in \mathbb{N}^{n \times m}$ denote the binary implicit feedback among n users and m items. Its element $x_{ij} \in \{0, 1\}$ indicates whether the j-th item is interacted by the i-th user. $\mathbf{x}_i = [x_{i1}, ..., x_{im}]^\top \in \mathbb{N}^m$ is a binary vector demonstrating the click history of user i on all items. The general goal of our method is to determine the feature representation $\mathbf{z}_i \in \mathbb{R}^d$ by defining some generative process from feature representation to preference feedback data,i.e., $p(\mathbf{x}_i|\mathbf{z}_i)$. The dimensionality d is usually fixed with a small value ($d \ll min\{n, m\}$).

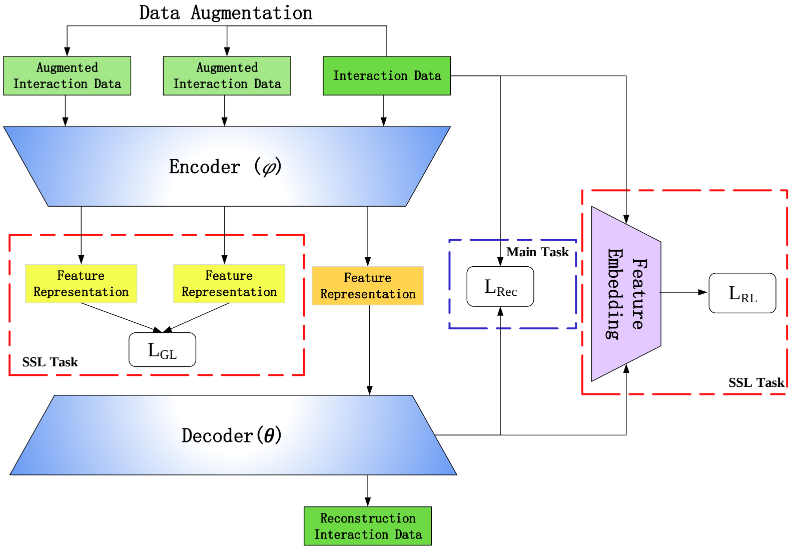


Fig. 1 Model architecture of the proposed DSVAE algorithm.

3.2 Overview

The overview architecture of the proposed DSVAE is illustrated in Fig. 1, which is accomplish by two tasks: main supervised task and dual self-supervised task. Specially, main supervised task employs variational autoencoder to generate the user feedback data. Dual SSL task contains two SSL task for improving both the generalization ability of VAE model on the sparse interaction datasets and personalized characteristic of recommendation results. Finally, our model jointly optimizes the objective of both recommendation task and dual SSL task.

3.3 Main task: Generating User Feedback Data

As depicted in Fig. 1, in order to predict user perference, we introduce variational autoencoder to generate user feedback data, which terms as main recommendation task. Specifically, the VAE module consists of two main components: Encoder and Decoder.

Encoder. The encoder model takes user feedback data $\mathbf{x}_i$ as input, and transforms it into a latent feature space by parametric update functions g_ϕ which can output parameters $\mu \in \mathbb{R}^d$ and $\theta \in \mathbb{R}^d$ for Gaussian distribution. In our framework, g_ϕ is designed as multi-layer perceptrons (MLPs), and the inference of $\mathbf{z}_i$ for the corresponding $\mathbf{x}_i$ is performed as:

$$q_\phi(\mathbf{z}_i|\mathbf{x}_i) = N(\mu, diag(\sigma^2)), \text{ with } \mu, \sigma^2 = g_\phi(\mathbf{x}_i) \tag{1}$$

The feature representation $\mathbf{z}_i$ is sampled from $q_\phi(\mathbf{z}_i|\mathbf{x}_i)$ and then passed to the subsequent decoder layer as input.

Decoder. In decoder layer, we suppose that reconstructed user feedback $\mathbf{x}_i$ is sampled from a generative process $p_\theta(\mathbf{x}_i|\mathbf{z}_i) = Mult(n, \pi(\mathbf{z}_i))$. The generator is a deterministic neural network $f_\theta(\mathbf{z}_i)$ parameterized by θ, the output of which is normalized via a *softmax* function to produce a probability vector $\pi(\mathbf{z}_i)$ over the entire item set.

$$\pi(\mathbf{z}_i) = softmax(f_\theta(\mathbf{z}_i)) \tag{2}$$

$$\mathbf{x}_i \sim Mult(n_i, \pi(\mathbf{z}_i)) \tag{3}$$

And the log-likelihood for $\mathbf{x}_i$ is:

$$\log p_\theta(\mathbf{x}_i|\mathbf{z}_i) = \sum_j x_{ij} \log \pi_j(\mathbf{z}_i) \tag{4}$$

All the above models can be trained by minimizing the negative of evidence lower bound (ELBO) in Eq. 5 below

$$\begin{aligned} \log p(\mathbf{x}_i; \theta) \geq & \mathbb{E}_{q_\phi(\mathbf{z}_i|\mathbf{x}_i)} [\log p_\theta (\mathbf{x}_i \mid \mathbf{z}_i)] \\ & - KL(q_\phi(\mathbf{z}_i \mid \mathbf{x}_i) \parallel P(\mathbf{z}_i)) \\ & \equiv \mathcal{L}_{main}(\mathbf{x}_i; \theta, \phi) \end{aligned} \tag{5}$$

where $P(\mathbf{z}_i)$ is the prior, which is assumed to be $N(0, \mathbf{I})$. The first term denotes the reconstruction loss between $\mathbf{x}_i$ and $\hat{\mathbf{x}}_i$, and the second term is the KL divergence, which prevents the conditional $q_\phi(\mathbf{z}_i \mid \mathbf{x}_i)$ from deviating from the Gaussian prior $N(0, \mathbf{I})$.

3.4 Dual Self-supervised Learning Task

As illustrated in Fig. 1, self-supervised learning is introduced to improve both the generalization ability of VAE model on the sparse interaction datasets and personalized characteristic of recommendation results. Thus, two SSL task are defined as follows:

Generalization ability enhancing self-supervised learning task. In order to enhance the generalization ability of VAE, we construct SSL task which aligns the representations learned from different views. The SSL task includes two procedures: data augmentation and contrastive learning, respectively.

(1) **Data Augmentation.** Given user history interaction, the key idea is to create two augmented examples by masking part of user interaction. With this operation, we term different augmented examples as different views

Algorithm 1 The Algorithm of **DSVAE**

```
Inputs: Implicit feedback matrix $\mathbf{X} \in \mathbb{N}^{n \times m}$, hyper-parameters $\lambda_1, \lambda_2, \tau_1, \tau_2$, max training epochs: B, batch size N.
Output: $\theta, \phi$
while epoch $\leq$ B, do
  //End-to-end optimization
  for sample mini-batch $\{\mathbf{x}_i\}_{i=1}^{n}$ , do
    for $\mathbf{x} \in \{1, 2, ..., n\}$, do
      //Construct data augmentations
      $\mathbf{x}_i' = \mathbf{P}' \odot \mathbf{x}_i$, $\mathbf{x}_i'' = \mathbf{P}'' \odot \mathbf{x}_i$
      //Feature representation learning via inference network
      $q_\phi(\mathbf{z}_i|\mathbf{x}_i) = N(\mu, diag(\sigma^2))$
      $q_\phi(\mathbf{z}_i'|\mathbf{x}_i') = N(\mu', diag(\sigma'^2))$
      $q_\phi(\mathbf{z}_i''|\mathbf{x}_i'') = N(\mu'', diag(\sigma''^2))$
      //generating reconstruction via generation network
      $\log p_\theta(\mathbf{x}_i|\mathbf{z}_i) = \sum_j x_{ij} \log \pi_j(\mathbf{z}_i)$
    end for
    //Multi-Task optimization
    $\mathcal{L} = \mathcal{L}_{main} + \lambda_1 \mathcal{L}_{gcl} + \lambda_1 \mathcal{L}_{rcl}$
    Update $\theta$ and $\phi$ of VAE to minimize $\mathcal{L}$;
  end for
end while
```

for the same user interaction, in which case our method can explore internal correlation between them. For each user feedback vector, we can randomly mask the value of positive items with a ratio α, which can be expressed as follows:

$$\mathbf{X}' = \mathbf{P}' \odot \mathbf{X}, \, \mathbf{X}'' = \mathbf{P}'' \odot \mathbf{X} \tag{6}$$

where $\mathbf{P}', \mathbf{P}'' \in \{\ 0, 1\}^{|m|}$ are independent masking vectors which are applied on feedback matrix $\mathbf{X}$ to generate two augmented views $\mathbf{X}'$ and $\mathbf{X}''$. As such, this augmentation operation is expected to learn internal relationship between $\mathbf{X}'$ and $\mathbf{X}''$.

(2) **Contrastive Learning.** With augmented views $\mathbf{X}'$ and $\mathbf{X}''$, we further employ contrastive learning to align the representations $\mathbf{z}_i'$ and $\mathbf{z}_i''$ which are learned from different views. Here, we treat $\{(\mathbf{x}_i^{'}, \mathbf{x}_i^{''}) \mid i \in \mathbf{U}\}$ as positive pair and $\{(\mathbf{x}_i^{'}, \mathbf{x}_j^{''}) \mid i, j \in \mathbf{U}, i \neq j\}$ as negative pair, where $\mathbf{U}$ is the set of users. To encourage the above properties, contrastive loss is represented as

$$\mathcal{L}_{gcl} = \sum_{i \in U} -\log \frac{\exp(s(\mathbf{z}_i^{'}, \mathbf{z}_i^{''})/\tau_1)}{\sum_{j \in \mathbf{U}} \exp(s(\mathbf{z}_i^{'}, \mathbf{z}_j^{''})/\tau_1)} \tag{7}$$

$$\mathbf{Z}' = g_\phi(\mathbf{X}'), \mathbf{Z}'' = g_\phi(\mathbf{X}'') \tag{8}$$

where g_ϕ is encoder network mentioned in Section 3.3 which serves to learn feature representations matrix $\mathbf{Z}'$ and $\mathbf{Z}''$. $s(\cdot)$ is cosine similarity, τ_1 is a tunable hyper-parameter for the softmax temperature. By minimizing contrastive loss, $\mathbf{z}_i'$ is enforce to be more similar to positive sample $\mathbf{z}_i''$ than other negative samples $\mathbf{z}_j'$.

Personalized characteristic enhancing self-supervised learning task. Unlike contrastive learning method mentioned above, which generates positive pair by data augmentation, the reconstruction and input matrix of VAE naturally fit into the contrastive term. To be specific, given user feedback matrix $\mathbf{X}$ and its reconstruction $\hat{\mathbf{X}} \in \mathbb{N}^{n \times m}$, we separately take $\{(\hat{\mathbf{x}}_i, \mathbf{x}_i) \mid i \in \mathbf{U}\}$ as positive pair and $\{(\hat{\mathbf{x}}_i, \mathbf{x}_j) \mid i, j \in \mathbf{U}, i \neq j\}$ as negative pair.

Then contrastive learning is adopted to optimize VAE by maximizing the agreement between user feedback vector and its own reconstruction. It encourages reconstruction to be as close as possible to the corresponding user feedback while being different from all rest user feedback. The contrastive loss is defined as,

$$\mathcal{L}_{rcl} = \sum_{i \in U} -\log \frac{\exp(s(h\left(\hat{\mathbf{x}}_i\right), h\left(\mathbf{x}_i\right)/\tau_2)}{\sum_{j \in \mathbf{N}_i} \exp(s(h\left(\hat{\mathbf{x}}_i\right), h\left(\mathbf{x}_j\right))/\tau_2)} \tag{9}$$

here $h(\cdot)$ is designed as multi-layer perceptrons to extra feature embedding. $\mathbf{N}_i$ is negative sample set, $s(\cdot)$ is cosine similarity and τ_2 is a tunable hyper-parameter for the softmax temperature.

As such, minimizing $\mathcal{L}_{rcl}$ can lead to more personalized recommendation result, i.e. distinguishing each reconstruction different from the others.

3.5 Optimizing Multi-task Learning

To enable SSL strategy learned representation to boost main recommendation performance, we optimize a combined objective of recommendation task and pretext task,

$$\mathcal{L} = \mathcal{L}_{main} + \lambda_1 \mathcal{L}_{gcl} + \lambda_2 \mathcal{L}_{rcl} \tag{10}$$

here $\mathcal{L}_{gcl}$ and $\mathcal{L}_{rcl}$ are defined respectively in Eq.(7) and Eq.(9); λ_1 and λ_2 are hyper-parameters to control the strength of contrastive SSL loss; $\mathcal{L}_{VAE}$ is defined in Eq.(5);

The algorithmic framework of our proposed DSVAE method is given in Algorithm 1. We randomly initialize parameters of encoder, decoder and feature embedding layer. A training epoch consists of two tasks: main recommendation task and dual SSL task.

4 Experiment

In this section, we evaluate the proposed DSVAE model on three datasets by comparing with state-of-the-art methods.

Table 1 Statistics of three datasets.

Datasets	**ML-1M**	**Douban Book**	**Epinions**
Users	6,040	13,024	18,088
Items	3,706	22,347	261,649
Ratings	1,000,209	792,062	764,352
Density	2.7%	0.255%	0.019%
User Information	Age, Gender, Occupation	$*$	$*$

4.1 Experimental Setup

To comprehensively and effectively evaluate the performance of our proposed DSVAE approach, we conduct experiments on three real-world datasets:

- ML-1M [5]: ML-1M collects user-movie interaction data from movie domain, and contains 6,040 users, 3,706 items and 1,000,209 ratings which are 10 discrete numbers within range of $[0.5, 5]$ and density of ratings is 2.7%. User content information is also included, i.e., age, gender, occupation.
- Douban Book [30]: Douban is the largest book review website in China, where user can rate or review books. It includes 13,024 users, 22,347 items and 792,062 ratings which are 5 discrete numbers within range of $[1, 5]$ and density of ratings is 0.255%.
- Epinions [17]: Epinions is extracted from consumer review website Epinions. It contains 18,088 users, 261,649 items and 764,352 ratings which are 5 discrete numbers within range of $[1, 5]$ and density of ratings is 0.019%.

Table 1 summarizes the characteristics of these employed experimental datasets. All the datasets are processed following [12]. We binarize explicit data by keeping ratings of four and higher and reserve users who have bought or rated at least five items.

Two widely used collaborative filtering evaluation metrics are adopted for performance comparisons, i.e., Recall and NDCG.

- Recall: Recall ratio of the ground truth item. Recall is used to test whether recommendation item is in the top-K list.

$$\mathbf{Recall}@k(i) = \frac{|\mathbf{R}_i \cap \mathbf{T}_i|}{|\mathbf{T}_i|} \tag{11}$$

  Here, $\mathbf{R}_i$ stands for the set of recommender items to user i. It's obtained by sorting likelihood of items which is predicted by decoder in VAE, excluding items from the training set. $\mathbf{T}_i$ denotes the set of favorite items of user u.
- NDCG: Normalized Discounted Cumulative Gain (NDCG) is adopted to measure the item ranking accuracy which assigns higher scores to top ranked items to consider the position of correctly recommended items.

$$\mathbf{NDCG}@k\,(i) = \frac{\mathbf{DCG}@k\,(i)}{\mathbf{IDCG}@k\,(i)} \tag{12}$$

Table 2 Performance comparison with several rating-only VAE based CF models.

Data	Metric	Mult-VAE	Improve	RecVAE	Improve	DSVAE
Douban Book	NDCG@25	0.1321	19.9%	0.1418	11.7%	**0.1584**
	NDCG@50	0.1530	16.1%	0.1633	8.8%	**0.1776**
	Recall@25	0.1785	8.7%	0.1771	9.5%	**0.1940**
	Recall@50	0.2383	4.7%	0.2418	3.2%	**0.2495**
ML-1M	NDCG@25	0.2977	17.8%	0.3224	8.8%	**0.3507**
	NDCG@50	0.3388	13.5%	0.3773	2.0%	**0.3847**
	Recall@25	0.3532	14.2%	0.3842	5.0%	**0.4032**
	Recall@50	0.4602	8.3%	0.4939	1.0%	**0.4982**
Epinions	NDCG@25	**0.0466**	-10.6%	0.0346	21.7%	0.0421
	NDCG@50	**0.0564**	-10.0%	0.0422	20.8%	0.0510
	Recall@25	**0.0845**	-11.7%	0.0648	16.7%	0.0756
	Recall@50	**0.1217**	-10.1%	0.0929	18.9%	0.1105

$$\mathbf{DCG}@k(i) = \sum_{n=1}^{k} \frac{2^{\mathbb{1}(\mathbf{R}_i \in \mathbf{T}_i)} - 1}{\log(\mathbf{n}+\mathbf{1})} \tag{13}$$

where IDCG is the DCG value with perfect ranking, $\mathbf{R}_i$ denotes the set of recommendation items to user i and $\mathbf{T}_i$ denotes the set of favorite items of user i, $\mathbb{1}$ is the indicator function.

We compare the proposed approach with four state-of-the-art methods, including two interaction-only VAE based recommendation approaches, and two content-aware VAE based recommendation approaches. Interaction-only VAE based recommendation approaches including Mult-VAE and RecVAE which are frequently used as comparison algorithms in VAE based CF methods. Content-aware VAE based recommendation approaches including Split-Merge CVAE and JVAE-CF which use side information to alleviate data sparsity problem. We demonstrate the effectiveness of SSL by comparing with content-aware VAE models.

For all comparison methods, we use the source code provided by the corresponding authors to reproduce experiment result or directly adopt best result presented in original paper.

- Mult-VAE [12]: A typical VAE based CF approach, which applies variational autoencoder model with multinomial likelihood to approximate recommendation ranking loss and improves the predictive performance. And it uses Bayesian inference for parameter estimation.
- Rec-VAE [22]: A complex variant of Mult-VAE, which proposes a new approach to set β hyper-parameter for Kullback-Leibler term in objective

function and replaces Gaussian distribution prior with composite distribution prior. The composite distribution prior is a mixture of standard Gaussian prior and feature representation distribution.
- Split-Merge CVAE [18]: An conditional variational autoencoder which concentrates on learning with label verification signals to ensure an exclusive latent mean factor for users with the same labels and leverages split-merge framework to handle complex multi-label combinations. Content information aims to promote model to distinguish and cluster users in latent subspace.
- JVAE-CF [11]: JVAE-CF encompasses VAE through augmenting structures to model both content information and user interaction information, which adds additional latent variable to extract high-level features associated with auxiliary information.

Moreover, to verify the effectiveness of each component used by DSVAE, a series of degenerate variants of DSVAE are also included for the comparisons:

- DSVAE-wGL: A degenerate variant of DSVAE. The difference between DSVAE-wGL and DSVAE is that the former does not use generated data to construct $\mathcal{L}_{gcl}$
- DSVAE-wRL: A weak version of DSVAE and an upgraded version of RecVAE, which omits $\mathcal{L}_{rcl}$ loss.

The parameters of all baselines are adopted from their original papers. As for DSVAE method, we take RecVAE as the VAE model and inherit the optimal values of hyper-parameters from RecVAE. For the unique part of DSVAE, we fine-tuned dual SSL task hyper-parameters to achieve best performance. In detail, the softmax temperature τ_1 and τ_2, the number of negative sample $\mathbf{N}_i$ and dropout ratio α is discussed in Section 4.5. DSVAE is implemented in PyTorch, the optimizer is Adam and hyper-parameters are automatically tuned according to the TPE method.

4.2 Performance comparison with several interaction-only VAE based CF models

We compare our DSVAE with two interaction-only VAE based CF methods (Mult-VAE, Rec-VAE), and the corresponding quantitative results in terms of NDCG and Recall are tabulated in Table 2, where the best performance is boldfaced and percentage indicates the relative improvement of our approach over other compared methods.

As shown in Table 2, our proposed model achieves 0.3847 in terms of NDCG@25, which is respectively 8.8% and 17.8% superior over RecVAE and Mult-VAE on ML-1M dataset. And our proposed model achieves 0.1418 in terms of NDCG@25 , which is respectively 19.9% and 11.7% superior over RecVAE and Mult-VAE on Douban Book dataset. However, the performance of RecVAE on Epinion datasets is worse than Mult-VAE so that DSVAE

Table 3 Performance comparison with content aware VAE based CF models on dataset ML-1M.

Model	JVAE-CF	Improve	Split-Merge CVAE	Improve	DSVAE
NDCG@25	0.2963	7.1%	0.2905	9.2%	**0.3171**
NDCG@50	0.2794	27.8%	0.2543	40.5%	**0.3572**
Recall@25	0.3423	4.5%	**0.3750**	-4.6%	0.3577
Recall@50	0.4256	9.7%	0.4500	3.8%	**0.4669**

is not as good as Mult-VAE. The experiments mentioned above powerfully demonstrate the effectiveness of DSVAE, and we attribute the success to the superiority of dual self-supervised task. Particularly, we observe that improvements on Douban Book and Epinions are more significant than ML-1M. This might be caused by data characteristics. Specifically, in Douban Book and Epinions, user-item interaction data is too sparse to guide the feature representation learning in VAE, while benefiting from the self-supervised learning task, DSVAE obtains auxiliary supervisions information to assist the feature representation learning.

4.3 Performance comparison with content aware VAE based CF models

Next, we compare our DSVAE with two content aware VAE based CF methods (Split-Merge CVAE, JVAE-CF) on ML-1M which contains user information. For fair comparison, we conduct DSVAE with the same backbone as Split-Merge CVAE and JVAE-CF. The experiment results are summarized in Table 3. We can find that DSVAE outperforms Split-Merge CVAE and JVAE-CF. It is worth noting that Split-Merge CVAE and JVAE-CF exploits both interaction data and content information, while DSVAE only uses interaction data. This observation demonstrates that DSVAE can add additional supervised information which is as useful as content information.

4.4 Ablation Experiments

In this section, we conduct an ablation study on DSVAE to further analyze the contribution of two SSL tasks. Specifically, we compare our proposed DSVAE method with RecVAE, DSVAE-wgcl and DSVAE-wrcl. The performance comparison is shown in Table 4. We have the following observations:

Effect of $\mathcal{L}_{gcl}$: To illustrate the effect of $\mathcal{L}_{gcl}$ in VAE based recommendation task, we first make comparison between DSVAE and DSVAE-wgcl on three real-world data sets. As shown in Table 4, DSVAE significantly outperforms DSVAE-wgcl on three datasets, which shows that our proposed $\mathcal{L}_{gcl}$ is effective for VAE based recommendation models. Exploring correlation between two fractions of the same user interaction is beneficial to learn a better feature representation. In addition, DSVAE-wrcl outperforms RecVAE even without

Table 4 Performance of our proposed DSVAE model with DSVAE-wRL,DSVAE-wCL and RecVAE with the evaluation metric NDCG@25,50, Recall@25,50 on ML-1M, Douban Book and Epinions Datasets.

Data	Metric	RecVAE	DSVAE-wrcl	DSVAE-wgcl	DSVAE
Douban Book	NDCG@25	0.1418	0.1516	0.1532	**0.1584**
	NDCG@50	0.1633	0.1772	0.1741	**0.1776**
	Recall@25	0.1771	0.1881	0.1855	**0.1940**
	Recall@50	0.2418	0.2476	0.2462	**0.2495**
ML-1M	NDCG@25	0.3224	0.3418	0.3465	**0.3507**
	NDCG@50	0.3773	0.3776	0.3834	**0.3847**
	Recall@25	0.3842	0.3979	0.3886	**0.4032**
	Recall@50	0.4939	0.5003	0.4911	**0.4982**
Epinions	NDCG@25	0.0346	0.0384	0.0378	**0.0421**
	NDCG@50	0.0422	0.0470	0.0474	**0.0510**
	Recall@25	0.0648	0.0672	0.0669	**0.0756**
	Recall@50	0.0929	0.0999	0.1037	**0.1105**

Table 5 The Optimal Value of τ_1, τ_2, α, $\mathbf{N}_i$ for DSVAE

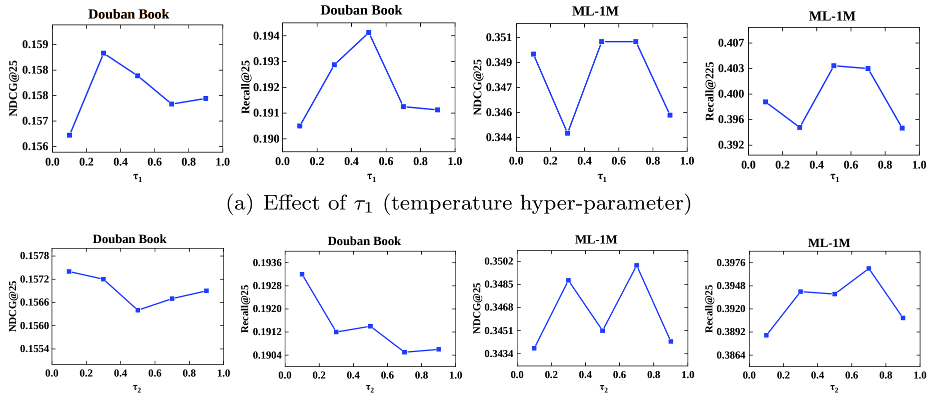
Parameters	**ML-1M**	**Douban Book**	**Epinions**
τ_1	0.5	0.5	0.05
τ_2	0.2	0.2	0.03
α	0.5	0.5	0.5
$\mathbf{N}_i$	80	128	150

$\mathcal{L}_{rcl}$. This observation illustrates the superiority of our $\mathcal{L}_{gcl}$ contrastive loss in VAE based recommendation models.

Effect of $\mathcal{L}_{rcl}$: To demonstrate the necessity of $\mathcal{L}_{rcl}$, we compare DSVAE with DSVAE-wRC and DSVAE-wgcl with RecVAE. From Table 4, we observe that DSVAE outperforms DSVAE-wrcl on three datasets, which verifies the effectiveness of $\mathcal{L}_{rcl}$. Besides, by comparing RecVAE with DSVAE-wgcl, we find the latter outperforms the former significantly, which also verifies the superiority of our proposed $\mathcal{L}_{rcl}$.

4.5 Impact of Hyper-parameters

Furthermore, we study the performance of our proposed method given different parameter settings. The parameter includes τ (temperature parameter in contrastive loss), α (dropout ratio in data augmentation), $\mathbf{N}_i$ (the number of negative samples). Fig. 2, Fig. 3 and Fig. 4 respectively illustrate how DSVAE performs under different τ, α and $\mathbf{N}_i$ configurations. We study the



(a) Effect of τ_1 (temperature hyper-parameter)

(b) Effect of τ_2 (temperature hyper-parameter)

Fig. 2 Impact of τ_1 and τ_2 on ML-1M, Douban Book

sensitivity analysis of DSVAE in the following subsections. Due to the space limitation, we omit the results on Epinions which have a similar trend to that on ML-1M and Douban Book. And the empirically optimal values of these hyper-parameters are shown in Table 5.

4.5.1 Effect of SSL weight τ

The temperature parameter τ_1 and τ_2 control penalties strength on hard negative samples [24]. To investigate the effect of τ empirically, we conduct the experiments under different τ configuration and express the comparing results in Fig. 2. As described in Fig. 2, with the increasing of τ_1 and τ_2, the performance of DSVAE at first increases and later decreases. And such phenomenon is intuitive, i.e., algorithm with smaller τ_1 and τ_2 indicates that too much attention to a few negative samples will loss the supremacy of adding multiple negative samples in the SSL objective; and larger τ_1 and τ_2 will lead model fall short in the ability to discriminate hard negatives samples.

4.5.2 Effect of SSL weight α

In this part, we discuss the effect of dropout ratio α. The results are presented in Fig. 3 in terms of NDCG@25 and Recall@25 on ML-1M and Douban Book datasets. As shown in Fig. 3, the best performance of ML-1M is obtained when α is 0.5 and best performance of Douban Book is obtained when α at around 0.5. We can observe that the performance curves of DSVAE trend to be worse when α exceeds a threshold. It is not hard to explain that, when the value of α is too small, the interaction has a lower possibility to be contained in augmented data so that SSL task contributes less or even damages the performance of the main task.

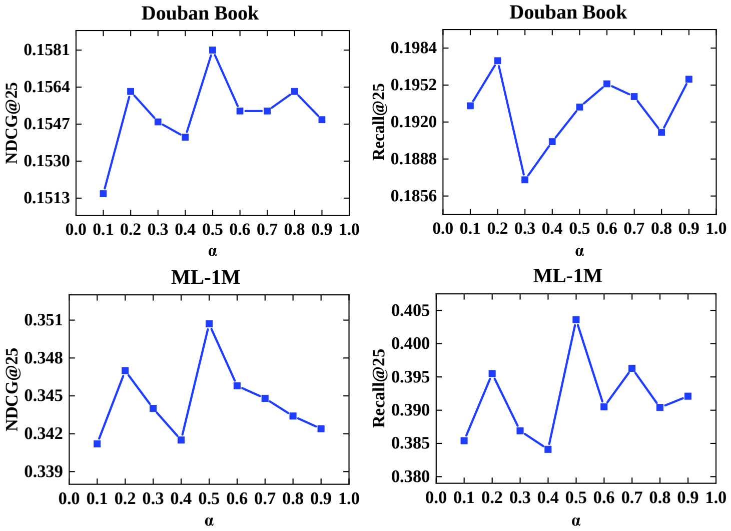


Fig. 3 Impact of α on ML-1M, Douban Book

4.5.3 Effect of negative samples $\mathbf{N}_i$

$\mathbf{N}_i$ controls sampling space of $\mathcal{L}_{rcl}$ on each user. We varied the amount of negative samples and summarized the performance in Fig. 4.

We find that best result is achieved with different value of $\mathbf{N}_i$ for two datasets, where a large $\mathbf{N}_i$ leads to a better performance in most cases. On the other hand, large $\mathbf{N}_i$ will draw more noise into the learning framework and small $\mathbf{N}_i$ will lose more valuable information, two of which have negative effects on the learning model.

5 Conclusion

In this paper, we have developed DSVAE, a new framework that integrates two self-supervised tasks into a single variational autoencoder framework. In sharp contrast to existing VAE based CF models, our DSVAE approach considers both intrinsical correlation and individual difference to improve both the generalization ability of VAE model on sparse interaction datasets and personalized characteristic of recommendation results. Extensive experiments demonstrated that our DSVAE approach achieves better performance than state-of-the-art methods. In the future, we will create more powerful self-supervised tasks, as well as research on a per-train model which utilizes self-supervised learning to obtain general user representation and then fine-tune top-K recommendation task with supervision signals.

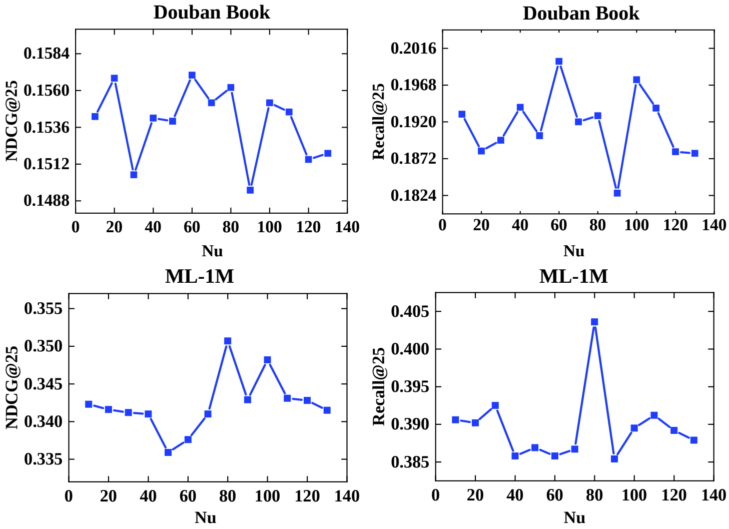


Fig. 4 Impact of $\mathbf{N}_i$ on ML-1M, Douban Book

References

1. Cacheda F, Carneiro V, Fernández D, Formoso V (2011) Comparison of collaborative filtering algorithms: Limitations of current techniques and proposals for scalable, high-performance recommender systems. ACM Trans Web
2. Devlin J, Chang M, Lee K, Toutanova K (2019) BERT: pre-training of deep bidirectional transformers for language understanding. In: NAACL-HLT, pp 4171–4186
3. Gopalan P, Hofman JM, Blei DM (2015) Scalable recommendation with hierarchical poisson factorization. In: UAI, pp 326–335
4. Gupta K, Raghuprasad MY, Kumar P (2018) A hybrid variational autoencoder for collaborative filtering. CoRR
5. Harper FM, Konstan JA (2016) The movielens datasets: History and context. ACM Trans Interact Intell Syst pp 19:1–19:19
6. He K, Fan H, Wu Y, Xie S, Girshick RB (2020) Momentum contrast for unsupervised visual representation learning. In: CVPR, pp 9726–9735
7. He X, Liao L, Zhang H, Nie L, Hu X, Chua T (2017) Neural collaborative filtering. In: WWW, pp 173–182
8. Hu Y, Koren Y, Volinsky C (2008) Collaborative filtering for implicit feedback datasets. In: ICDM
9. Kingma DP, Welling M (2014) Auto-encoding variational bayes. In: ICLR
10. Koren Y, Bell RM, Volinsky C (2009) Matrix factorization techniques for recommender systems. Computer
11. Lee W, Song K, Moon I (2017) Augmented variational autoencoders for collaborative filtering with auxiliary information. In: CIKM, pp 1139–1148

12. Liang D, Krishnan RG, Hoffman MD, Jebara T (2018) Variational autoencoders for collaborative filtering. In: WWW, pp 689–698
13. Liu H, Wen J, Jing L, Yu J (2019) Deep generative ranking for personalized recommendation. In: RecSys, pp 34–42
14. Liu H, Jing L, Wen J, Wu Z, Sun X, Wang J, Xiao L, Yu J (2020) Deep global and local generative model for recommendation. In: WWW, pp 551–561
15. Lobel S, Li C, Gao J, Carin L (2020) Ract: Toward amortized ranking-critical training for collaborative filtering. In: ICLR
16. Ma J, Zhou C, Cui P, Yang H, Zhu W (2019) Learning disentangled representations for recommendation. In: NeurIPS, pp 5712–5723
17. Massa P, Avesani P (2007) Trust-aware recommender systems. In: RecSys, pp 17–24
18. Pang B, Yang M, Wang C (2019) A novel top-n recommendation approach based on conditional variational auto-encoder. In: PAKDD, pp 357–368
19. Rakesh V, Wang S, Shu K, Liu H (2019) Linked variational autoencoders for inferring substitutable and supplementary items. In: WSDM, pp 438–446
20. Rendle S, Freudenthaler C, Gantner Z, Schmidt-Thieme L (2009) BPR: bayesian personalized ranking from implicit feedback. In: UAI, pp 452–461
21. Salakhutdinov R, Mnih A (2007) Probabilistic matrix factorization. In: Advances in Neural Information Processing, pp 1257–1264
22. Shenbin I, Alekseev A, Tutubalina E, Malykh V, Nikolenko SI (2020) Recvae: A new variational autoencoder for top-n recommendations with implicit feedback. In: WSDM, pp 528–536
23. Truong Q, Salah A, Lauw HW (2021) Bilateral variational autoencoder for collaborative filtering. In: WSDM, pp 292–300
24. Wang F, Liu H (2020) Understanding the behaviour of contrastive loss. CoRR
25. Wu J, Wang X, Feng F, He X, Chen L, Lian J, Xie X (2021) Self-supervised graph learning for recommendation. In: SIGIR, pp 726–735
26. Wu Y, DuBois C, Zheng AX, Ester M (2016) Collaborative denoising auto-encoders for top-n recommender systems. In: WSDM, pp 153–162
27. Wu Y, Macdonald C, Ounis I (2020) A hybrid conditional variational autoencoder model for personalised top-n recommendation. In: ICTIR '20, pp 89–96
28. Yao T, Yi X, Cheng DZ, Yu FX, Menon AK, Hong L, Chi EH, Tjoa S, Kang J, Ettinger E (2020) Self-supervised learning for deep models in recommendations. CoRR
29. Zhang Y, Zhu Z, He Y, Caverlee J (2020) Content-collaborative disentanglement representation learning for enhanced recommendation. In: RecSys 2020: Fourteenth ACM Conference on Recommender Systems, Virtual Event, Brazil, September 22-26, 2020, pp 43–52
30. Zhao G, Qian X, Xie X (2016) User-service rating prediction by exploring social users' rating behaviors. IEEE Trans Multim pp 496–506

31. Zhu Y, Min MR, Kadav A, Graf HP (2020) S3VAE: self-supervised sequential VAE for representation disentanglement and data generation. In: CVPR, pp 6537–6546