

Multi-Scale Distribution Deep Variational Autoencoder for Explanation Generation

ZeFeng Cai¹, Linlin Wang^{1*}, Gerard de Melo², Fei Sun³, Liang He¹

¹ East China Normal University

² Hasso Plattner Institute, University of Potsdam

³ Alibaba Group

oklen@stu.ecnu.edu.cn, {llwang,lhe}@cs.ecnu.edu.cn, gdm@demelo.org

ofey.sunfei@gmail.com

Abstract

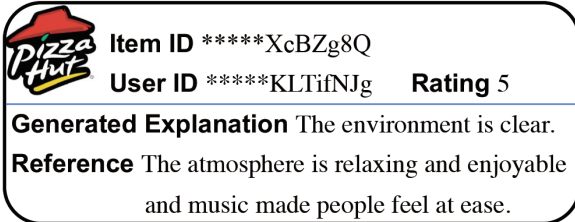
Generating explanations for recommender systems is essential for improving their transparency, as users often wish to understand the reason for receiving a specified recommendation. Previous methods mainly focus on improving the generation quality, but often produce generic explanations that fail to incorporate specific details of user and item. To resolve this problem, we present Multi-Scale Distribution Deep Variational Autoencoders (MVAE). A deep hierarchical VAE with a prior network that eliminates noise while retaining meaningful signals in the input, coupled with a recognition network serving as the source of information to guide the learning of the prior network. Further, the Multi-scale distribution Learning Framework (MLF) along with a Target Tracking Kullback-Leibler divergence (TKL) mechanism are proposed to employ multiple KL divergences at different scales for more effective learning. Extensive empirical experiments demonstrate that our methods can generate explanations with concrete input-specific contents.

1 Introduction

Due to the massive demand for convincing high-quality recommendations, researchers from both academic and industrial communities have paid increasing attention to the topic of enhancing the explainability of recommender systems (Wang et al., 2018b,a; Xian et al., 2019; Chen et al., 2019). Explanations for recommendations in real-world scenarios are presented in a variety of different forms, among them, the most popular and natural form is that of free-text explanations given in natural language (Zhang and Chen, 2020).

As shown in Fig. 1, this task requires a machine to generate a textual explanation based on a given user ID, item ID, and the rating score from a recommender system. Previous models attempt to

* Corresponding author. Email: llwang@cs.ecnu.edu.cn.




	Item ID *****XcBZg8Q
	User ID *****KLTifNJg Rating 5
Generated Explanation	The environment is clear.
Reference	The atmosphere is relaxing and enjoyable and music made people feel at ease.

Figure 1: An example of explanation generation.

embed these IDs in a similar way as normal words. However, since the IDs appear far less frequently than the words, most approaches typically fail to account for specific features of the users and item. Hence, it is a very common phenomenon to obtain explanations without concrete characteristics about the given user and item as shown in Table 4. A probable reason for this phenomenon is that these models fail to utilize the input embeddings effectively. Specifically, in most models, the user and item information is merely provided as randomly initialized input embeddings, which barely contain meaningful information, but introduce noise that may be indistinguishable from more meaningful information. Here, we refer to noise from the similarities of randomly initialized input embeddings that are conflated with implicit patterns contained in our data. For example, there may be two user embedding similar to each other while in our data they represent users very different from each other. Importantly, as the recommendation data is sparse, some of the noisy embeddings are not able to be adequately trained, resulting in that the noise dominates the representation of those embeddings, as shown in Section 4.5. Since the presence of noise disturbs the model’s ability to interpret the input embeddings at the inception of training, the model may tend to generate explanations in an unconditional manner. Moreover, such noisy inputs may still exist even after training. A common phenomena is that some users or items have very limited

relevant training instances. Consequently, their corresponding representation embeddings are insufficiently trained and remain noisy. Therefore, it is vital to overcome such noise, so as to ensure the model can generate in a conditional manner.

To deal with this problem, we present **Multi-Scale Distribution Deep Variational Autoencoders (MVAE)**. They consist of three modules, namely a recognition network, prior network, and a reconstruction network. The prior network in our model can filter out the noise contained in input embeddings, while retaining meaningful information for generation through information compression. Moreover, to help the prior network learn to generate fine-grained information, the recognition network is leveraged to provide the prior network with suitable guiding information. Thus, the decoder tends to generate explanations in a conditional manner with a substantially more informative generation signal.

However, with strong guiding signals available during training, generation becomes much simpler, which may result in a degradation of performance when such information is no longer available during testing. Thus, we propose a **Multi-scale distribution Learning Framework (MLF)** along with a target **Tracking Kullback-Leibler divergence (TKL)** mechanism to reduce this performance gap between training and testing. The optimization effectiveness of the prior network can further be boosted when this method is employed at multiple different scales.

Overall, our contributions are as follows:

- We highlight the problem of noise in the input embeddings that current approaches suffer from. To the best of our knowledge, MVAE is the first model that aims to overcome such noisy input embeddings in explanation generation for recommender systems.
- We propose MVAE, a novel VAE model for explanation generation, which can utilize the input embedding effectively for generating high-quality explanations. The prior network in our model filters the noise contained in the input embeddings, while retaining meaningful information for generation. Moreover, we propose multi-scale distribution learning framework along with a target tracking Kullback–Leibler divergence mechanism to improve the optimization of the prior network, yielding better generalization performance.
- Extensive experiments show that our approach yields state-of-the-art results on three real-world datasets, demonstrating its effectiveness in generating high-quality explanations. A series of in-depth analyses shed further light on its ability to overcome noise contained in input embeddings in the training process.

2 Related Work

For generation of textual explanations, mainstream research can be divided into two categories: template-based and natural language generation approaches. Template-based approaches generate explanations by filling the slots of predefined templates (Zhang et al., 2014), which are typically manually specified in advance. Natural language generation approaches, in contrast, adopt an encoder–decoder framework such as a recurrent seq-to-seq model (Li et al., 2020) or a Transformer-based architecture (Li et al., 2021) to learn to generate more diverse explanations based on the respective input.

In recent years, the latter strategy has received considerable attention, mainly owing to advances in neural generation along with the massive availability of text from online review systems.

Still, existing natural language generation methods may generate overly generic sentences that fall short at providing concrete information and are thus less useful for users (Cao et al., 2018). Indeed, explanation generation goes beyond mere generation, as it is expected to improve the transparency of the recommendation engine (Tintarev and Masthoff, 2015). Thus, technical ideas to encourage the generation process to account for more conditional signals are crucial to enable models to generate more specific explanations that are custom-tailored for particular user–item pairs.

Variational autoencoders (VAE) were proposed by Kingma and Welling (2014) based on the idea of autoencoding, which has been used for noise reduction (Vincent et al., 2008, 2010). VAEs have been studied extensively in a variety of language generation tasks, including text summarization (Li et al., 2017a) and dialogue generation (Serban et al., 2017; Wen et al., 2017; Zhao et al., 2017). A VAE maximizes the mutual information between the input and latent variables (Barber and Agakov, 2003; Alemi et al., 2017), requiring the network to retain the information content of the input data to the extent possible (Shwartz-Ziv and Tishby, 2017). Hence, VAEs are qualified to overcome

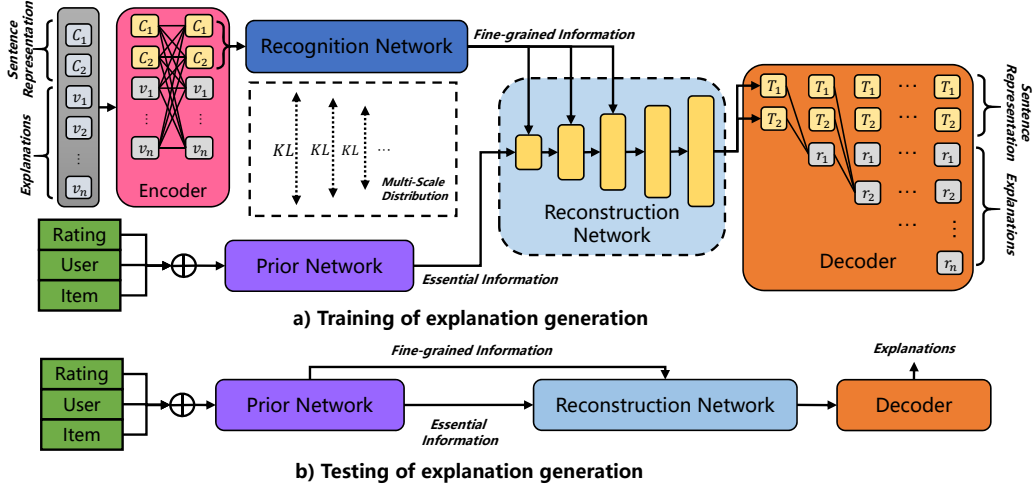


Figure 2: Overview of the Proposed Model.

the overly generic explanations caused by uninformative noisy input embeddings and prompt the construction of more meaningful outputs.

3 Proposed Model

An overview of our model is given in Fig. 2. The recognition network encodes the explanations and generates fine-grained information for the reconstruction network. The prior network encodes the input embeddings and generates essential information for the reconstruction network. The essential information here refers to the general semantics of a reason, which can be described in multiple ways, while the fine-grained information here refers to information that determines the details in the explanations, thus narrowing down and customizing the essential information to a specific form.

Finally, the reconstruction component decodes the given information and generates explanations. Additionally, the proposed MLF employs KL divergence at multiple different scales, which improves the optimization of the prior network. The TKL applied in every KL divergence can aid the learning of the prior network even further. We will present the details of each network in the following sections.

3.1 Input Encoding

To achieve a suitable transformation for compression and reconstruction of information, we design a basic component called the representation transformation module, which is used repeatedly in our

model. Formally, it can be defined as follows:

$$\begin{aligned}
 f_{d_x, d_y}(x) &= \text{SN}(W_{d_x \times d_y} \text{GELU}(x) + b_{d_x}) \\
 T_{d_1, d_2, d_3}(x) &= f_{d_2, d_3} \circ f_{d_2, d_2} \circ f_{d_1, d_2} \circ f_{d_1, d_1} \\
 x' &= \text{LayerNorm}(T_{d_1, d_2, d_3}(x) + x) \\
 y &= f_{d_3, d_4}(x')
 \end{aligned} \tag{1}$$

Here, $x \in \mathbb{R}^{d_x}$ is the input and $y \in \mathbb{R}^{d_y}$ is the output of this module. The subscripts d_x, d_y of f and d_1, d_2, d_3 of any F are the dimensionalities of the matrices or vectors used in the corresponding function. T is a composite module consisting of four different f , where \circ denotes composition, SN is the spectral normalization introduced by Yoshida & Miyato (2017). GELU (Hendrycks and Gimpel, 2016) is an activation function based on the cumulative distribution function for a Gaussian Distribution.

For simplicity, we denote this module as $\text{Block}(\cdot)$. Moreover, our notation assumes that its output is split into equal-sized partitions if the output is assigned to more than one variable.

Recognition Network The recognition network serves to provide guidance to the prior network to enable it better generate fine-grained information, while supplying fine-grained information to the reconstruction network in training, as shown in Fig. 2(a). With the ground-truth explanations as input, the recognition component can generate valuable guiding information.

We first employ Transformer (Vaswani et al., 2017) encoder layers to encode input tokens $v_i \in \mathbb{R}^{d_v}$ into compact hidden states. The two special tokens C_1 and C_2 represent the overall input. The

encoders are represented by B_b and the encoding process can be described as follows:

$$O_1, O_2, \dots, O_{n+2} = B_b(C_1, C_2, v_1, \dots, v_n) \quad (2)$$

Here, O_i is the i -th output of B_b . We concatenate O_1 and O_2 as the initial sentence-level representation $C'_0 = [O_1, O_2]$. Then the input information is compressed and the distributions of fine-grained information can be obtained as follows:

$$\begin{aligned} C'_i &= \text{Block}_{d_i}^R(C'_{i-1}) \\ \mu_{rz_j}, \sigma_{rz_j}, C'_j &= \text{Block}_{s_j}^R(C'_{j-1}) \end{aligned} \quad (3)$$

Here, $i \in \{1, 2, \dots, n_{rd}\}$, $j \in \{n_{rd} + 1, \dots, n_{rd} + n_{rs}\}$, while n_{rd} and n_{rs} are the number of Block_d^R and Block_s^R instances in the recognition network, respectively. Further, $\mu_{rz_j} \in \mathbb{R}^{d_{z_j}}$ is the mean and $\sigma_{rz_j} \in \mathbb{R}^{d_{z_j}}$ is the variance of the posterior distribution $q_{\theta_j}(z|x)$, where θ denotes the parameters of the recognition network. The reparameterization trick (Kingma and Welling, 2014) is used to sample a rz_j from $q_{\theta_j}(z|x)$.

Prior Network As for the prior network, its key aim is to filter out uninformative noise in the given input embeddings while retaining the essential signals for later reconstruction. The given user ID, item ID and rating are first mapped to their representation embeddings E_u, E_i, E_r and are then concatenated. After that, we employ a compression block Block_d^P to filter out noise in the input and an additional Block_s^P to generate fine-grained information:

$$\begin{aligned} E'_0 &= [E_u, E_i, E_r] \\ E'_i &= \text{Block}_{d_i}^P(E'_{i-1}) \\ \mu_{pz_j}, \sigma_{pz_j}, E'_j &= \text{Block}_{s_j}^P(E'_{j-1}) \end{aligned} \quad (4)$$

Here, $i \in \{1, 2, \dots, n_{pd}\}$, $j \in \{n_{pd} + 1, \dots, n_{pd} + n_{ps}\}$, while n_{pd} , n_{ps} refer to the number of Block_d^P and Block_s^P instances in the recognition network, respectively. Further, $\mu_{pz_j} \in \mathbb{R}^{d_{z_j}}$ and $\sigma_{pz_j} \in \mathbb{R}^{d_{z_j}}$ are the mean and variance of $q_{\phi_j}(z|E')$, where ϕ denotes the parameters of the prior network.

After suitable training, the prior network will be able to replace the recognition network to supply fine-grained signals to the reconstruction network in the testing phrase, as illustrated in Fig. 2(b).

3.2 Multi-Scale Learning

In our model, it is crucial to ensure that the prior network can learn suitable fine-grained information

at different scales from the recognition network effectively. To this end, we further propose the MLF and TKL techniques.

Target Tracking KL Regularizations (TKL)

Our TKL mechanism serves to improve the representation of the output latent variable z with regard to fine-grained information and thus ease the difficulty of learning a prior network for generation of specific fine-grained information. For simplicity, the subscripts to represent the index of layers are omitted here, but this mechanism is applied to every pair of distributions of prior network and recognition network with the same input variable scale. The TKL consists of two KL divergences: the first is $\text{KL}(q_{\theta}(z|x) \| q_{\phi}(z|E'))$ and the second is $\text{KL}(\mathcal{N}(0, I_{d_z}) \| q_{\theta}(z|x))$. Here, I_{d_z} denotes a diagonal matrix. Traditionally, VAE models directly apply KL divergence $\text{KL}(p(z|x) \| \mathcal{N}(0, I))$ on the final posterior distribution ($q_{\phi}(z|E')$ in our model), which is not suitable for our case, as the distribution $q_{\phi}(z|E')$ is learnt with $q_{\theta}(z|x)$ during the training phase. If we directly apply KL regularization between $\mathcal{N}(0, I_{d_z})$ and $q_{\phi}(z|E')$, the lagging problem (He et al., 2019) would cause posterior collapse. To resolve this problem, we use $\text{KL}(\mathcal{N}(0, I_{d_z}) \| q_{\theta}(z|x))$ to improve the quality of representation of latent variables z , as we find if both $\text{KL}(q_{\theta}(z|x) \| q_{\phi}(z|E'))$ and $\text{KL}(\mathcal{N}(0, I_{d_z}) \| q_{\theta}(z|x))$ are small enough, we can then obtain a small $\text{KL}(\mathcal{N}(0, I_{d_z}) \| q_{\phi}(z|E'))$. Finally, we can obtain:

$$\text{KL}(\mathcal{N}(0, I_{d_z}) \| q_{\phi}(z|E')) \approx \text{KL}(\mathcal{N}(0, I_{d_z}) \| q_{\theta}(z|x)) \quad (5)$$

Therefore, the first KL divergence term supports the second KL divergence term to implicitly apply disentangled regularization to improve the representation of fine-grained cues (Shao et al., 2020). Overall, the TKL mechanism applied to pairs of distributions can be expressed as

$$\begin{aligned} \text{TKL}(\mathcal{N}(\mu_{rz}, \sigma_{rz}) \| \mathcal{N}(\mu_{pz}, \sigma_{pz})) &= \\ \beta \text{KL}(\mathcal{N}(\mu_{rz}, \sigma_{rz}) \| \mathcal{N}(0, I_{d_z})) &+ \\ + \text{KL}(\mathcal{N}(\mu_{rz}, \sigma_{rz}) \| \mathcal{N}(\mu_{pz}, \sigma_{pz})), & \end{aligned} \quad (6)$$

where β is a hyperparameter originally from β -VAE (Higgins et al., 2017) to balance between reconstruction and disentangled regularization.

Multi-Scale Learning Framework (MLF)

The multi-scale distributions are originally proposed by Sønderby et al. (2016) to improve the flexibility of prior distribution and thus improve the generation

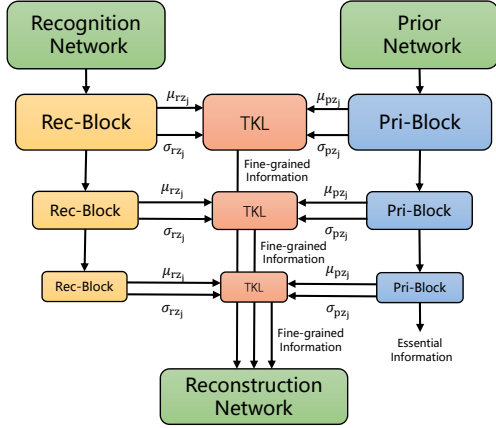


Figure 3: Multi-Scale Learning Framework. The Rec-Block represents the Block_s^R in the recognition network and Pri-Block represents the Block_s^P in the prior network.

quality of a VAE. We extend this architecture and the overall structure is shown in Fig. 3. Our MLF can also improve the flexibility of prior distributions and controls the fine-grained information to aid the reconstruction network. During training, rz_j from the recognition network is provided to the reconstruction network, delivering fine-grained information to assist the latter in achieving the reconstruction. During testing, the μ_{pz_j} from the prior network come into play. For simplicity and consistency, we refer to both with the same symbol z_j in the following.

More importantly, MLF decides how the prior block network is optimized according to the recognition network. Since multi-scale information is leveraged, the prior network can be better optimized. The sampling process from the distributions of the recognition network add appropriate noise into the supplementary information during training, which improves the denoising ability of the reconstruction network. Therefore, when the μ_{pz} without sampling noise but with noise from the input signals are used in testing, the reconstruction network can better cope with the situation of noisy supplementary information. This results in a reduction of the performance gap between training and testing. The overall regularization loss can be represented as:

$$\mathcal{L}_{\text{MLF}} = \sum_{n_{pd}+1}^{n_{ps}} \text{TKL}(\mathcal{N}(\mu_{rz}, \sigma_{rz})_j \parallel \mathcal{N}(\mu_{pz}, \sigma_{pz})_j) \quad (7)$$

3.3 Reconstruction Network

Reconstruction Network The reconstruction network is responsible for explanation generation according to received fine-grained information and essential information. The mechanism of the reconstruction network can be described as follows:

$$\begin{aligned} H'_0 &= E'_{n_{pd}+n_{ps}} \\ H'_j &= \text{Block}_j^D(H'_{j-1} + z_k) \\ H'_i &= \text{Block}_i^D(H'_{i-1}) \\ T_1, T_2 &= \text{chunk}(H_{n_{ps}+n_{pd}}) \end{aligned} \quad (8)$$

where $j \in \{1, \dots, n_{ps}\}$, $i \in \{n_{ps} + 1, n_{ps} + n_{pd}\}$, $k = n_{ps} + n_{pd} + 1 - j$. Block_*^D are used to reconstruct the information. The sentence representations $T \in \mathbb{R}^{d_v}$ are fed into a GPT decoder (Floridi and Chiriatti, 2020) as initial tokens. $\text{chunk}(\cdot)$ denotes splitting the input into two equal-sized parts.

The negative log-likelihood function is used as the objective function, which can be expressed as

$$\mathcal{L}_{\text{rec}} = - \sum_{t=1}^n \log(p(r_t^*)), \quad (9)$$

where r_t^* is the ground-truth review word at step t and n is the total length of the output token sequence.

3.4 Overall Objective Function

Ultimately, the optimization of our model is achieved using the following overall objective function:

$$\mathcal{L} = \mathcal{L}_{\text{rec}} + \mathcal{L}_{\text{MLF}} \quad (10)$$

4 Experiments

4.1 Dataset

For the evaluation, we use three large-scale datasets, including Yelp¹ for restaurants, Amazon 5-core Movie & TV² for movies, and TripAdvisor³ for hotels. In contrast to prior work, we only select and use challenging samples where related users or items have fewer than 15 reviews for Yelp and TripAdvisor, 20 reviews for Amazon movies. Our setting is suitable for advancing the research on this task. The statistics of the resulting **Yelp**, **Amazon**, and **TripAdvisor** datasets are given in Table 1.

¹<http://www.yelp.com/dataset>

²<http://www.jmcauley.uscd.edu/data/amazon>

³<http://www.tripadvisor.com>

Entries	<i>Amazon</i>	<i>Yelp</i>	<i>TripAdvisor</i>
# of users	161,434	451,937	333,409
# of items	118,862	154,951	304,954
# of reviews	653,568	1,033,823	1,311,676
Avg. # of reviews/user	4.04	2.28	3.93
Avg. # of reviews/item	5.49	6.67	4.30
Avg. # of words/explanation	14.81	15.03	14.84

Table 1: Statistics of three processed datasets.

4.2 Evaluation Metrics

We employ five metrics to evaluate the quality of generated explanations, including BLEU-1, BLEU-4, ROUGE-1, ROUGE-L, and METEOR. BLEU-1 and BLEU-4 are BLEU (Papineni et al., 2002) scores with 1-grams and 4-grams, respectively. ROUGE-1 refers to ROUGE (Lin, 2004) scores with 1-grams, while ROUGE-L finds the longest common subsequence and takes the sentence level structure similarity into account. METEOR (Banerjee and Lavie, 2005; Sharma et al., 2017) is a metric that correlates better at the sentence level with human evaluations. For all metrics, higher scores indicate better results.

4.3 Baselines

Various recent approaches serve as strong baselines in our experiments⁴. In addition, we consider several variants of our model to ascertain the effectiveness of our proposed techniques.

NRT (Li et al., 2017b): In this model, a multi-layer perceptron (MLP) is used to predict a rating based on the given user ID and item ID. It formulates the explanation generation task as a text summarization task and trains in a multi-task learning framework. In our case, the explanation sentence is used as the tip.

Att2Seq (Dong et al., 2017): This model employs a MLP to encode three attributes and adopts a two-layer LSTM to decode representations for generating textual explanations.

NETE (Li et al., 2020): A neural template explanation generation framework design with a gated fusion recurrent unit (GFRU) to generate neural templates and explanations in parallel. It combines advantages of both templates and neural networks.

PETER (Li et al., 2021): PETER is a Transformer-based model that reforms the atten-

⁴Note that our model can be adapted to arbitrary recommender systems, while some explainable recommendation baselines require access to specific internal information of the recommender system and are thus omitted for a fair comparison.

tion mask to combine different kinds of input embedding and finally be able to generate natural language explanations, which resulted in the previous state-of-the-art.

MVAE-NoKL: The second KL divergence regularization in TKL is removed, in order to investigate whether TKL can effectively apply disentangled regularization to latent variables for helping the reconstructing network to decode latent variable and easing the difficulty with which the prior network learns from the recognition network.

MVAE-NoMLF: In this variant, distributions of all scales of MLF are removed except for the smallest one. This allows us to investigate whether MLF can promote the learning of the prior network and supply suitable amounts of fine-grained information to the reconstruction network.

4.4 Implementation Details

Following common practice in recommender systems, we map a rating greater than or equal to 3 to positive sentiment, and consider it a negative sentiment otherwise. The final results are the average of 5 experiments with different random data splits. In the training phase, if the decrease ratio of the validation loss is larger than 0.98, we decrease the learning rate by a factor of 0.8. We set the longest generation length to 20, while the average length of sentences is about 15. For all of the models, we set a fixed vocabulary size of 20,000. For the hyperparameters of other models in the experiments, we adopt the default settings in their published code to ensure the proper performance.

For our model, we set the hidden sizes of the Transformer encoder and decoder layers to 768 and each consist of two layers. For the prior and recognition networks, we stack 6 Block units to compress the input by a factor of 0.5 in each Block. Another 6 layers of Block units are stacked for reconstruction in the reconstruction network. We use AdamW optimization (Kingma and Ba, 2015).

The β used in our TKL is set to 0.001 with the following annealing schedule:

$$\beta' = \beta \cdot \frac{1}{1 + \exp(-k(n_{\text{step}} - a_0))} \quad (11)$$

To select suitable hyperparameters for the annealing function, we first disable the second KL regularization and record how many steps our model needs to reach convergence. Then half of this amount of steps is chosen as a_0 . The weight $k = 0.0025$

	BLEU (%)		ROUGE-1 (%)			ROUGE-L(%)			METEOR(%)
	BLEU-1	BLEU-4	Precision	Recall	F1	Precision	Recall	F1	METEOR
<i>Yelp</i>									
NRT	5.90	0.41	7.36	5.71	6.43	5.51	4.68	5.06	2.43
Att2Seq	11.95	0.83	14.90	11.56	13.02	11.17	9.48	10.25	4.92
NETE	14.76	1.02	18.40	14.27	16.07	13.79	11.70	12.66	6.08
PETER	16.58	1.15	20.67	16.03	18.06	15.49	13.15	14.22	6.83
MVAE	21.42	2.25	21.07	16.93	18.77	17.17	13.76	15.28	7.26
Improvement (%)	29.19	95.91	1.94	5.61	3.98	10.85	4.64	7.40	6.30
<i>Amazon</i>									
NRT	5.61	0.39	6.99	5.42	6.11	5.24	4.45	4.81	2.31
Att2Seq	11.35	0.79	14.16	10.98	12.37	10.61	9.01	9.74	4.68
NETE	14.02	0.97	17.48	13.55	15.27	13.10	11.12	12.03	5.77
PETER	15.75	1.09	19.64	15.23	17.15	14.72	12.49	13.51	6.49
MVAE	19.35	2.10	20.12	15.98	17.81	16.71	13.27	14.79	7.24
Improvement (%)	22.84	92.70	2.44	4.96	3.84	13.56	6.24	9.48	11.61
<i>TripAdvisor</i>									
NRT	7.08	0.49	8.83	6.86	7.71	6.62	5.62	6.08	2.92
Att2Seq	14.34	0.99	17.88	13.87	15.62	13.40	11.38	12.31	5.91
NETE	17.71	1.23	22.08	17.12	19.28	16.54	14.04	15.19	7.29
PETER	19.90	1.38	24.90	19.24	21.67	18.59	15.78	17.07	8.20
MVAE	23.70	2.94	25.18	20.62	22.67	19.97	16.51	18.08	10.03
Improvement (%)	19.14	113.32	1.53	7.17	4.63	7.46	4.64	5.91	22.40

Table 2: Performance comparison of explanations generation of different methods on three datasets. Improvements are computed as relative gains compared with the previous state-of-the-art method. Best results are highlighted in boldface, and the statistical significance over the best baseline is $p < 0.05$ via a t -test.

is selected without any tuning. The learning rate warm-up step count is set to 5,000 for all datasets.

In training phase, the teacher-force strategy is employed for the decoder network to accelerate the training. The dropout rate used in the encoder network and decoder network is set to 0.3 and gradient clipping is applied with 5.0. For the multi-scale learning framework, n_{rd} is equal to n_{pd} and n_{rs} is equal to n_{ps} . The n_{rd} is set to 4 and n_{rs} is set to 3. In both the prior network and recognition network, the variable is compressed by the ratio of 0.5. In our model, the dimensionality of the input variable is 1,536 and the dimensionality of resulting encoding is 12 after 7-fold compression. Similarly, in the reconstruction network, the latent variable is reconstructed from size 16 to size 1,536 after 7 reconstruction blocks. In addition, the word embedding used in the encoder Transformer layers and decoder Transformer layers are shared.

4.5 Existence of Initial Noise

To show the existence of initial noise, we first conduct an additional experiment on the *Yelp* dataset.

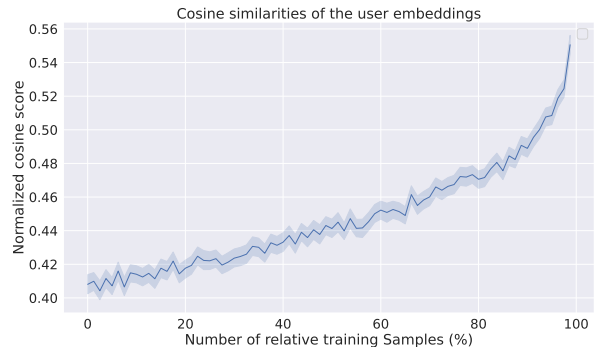


Figure 4: Illustration of the existence of initial noise.

Specifically, we randomly sample half of all examples, then duplicate them and all involved input embeddings to build a new dataset. In this dataset, there are two different instances of each user with their corresponding respective examples. Subsequently, we train a naive VAE model on the dataset. We sorted the user embeddings based on the number of their relevant training examples and calculate normalized cosine similarity between the two instances of the same user. We cluster them

into 80 bins to enable a clearer presentation of the extensive data. The results are shown in Fig. 4. Intuitively, the difference of two instances of the same user represents the noise contained in the embeddings, and we can see that as increasing the number of relevant training samples, the noise becomes smaller and smaller. We believe that this is because user embeddings with more training samples are updated more frequently, while we can see there is still substantial noise remaining on the embeddings with few relevant training samples. This motivates the necessity of employing our model to eliminate such noise.

4.6 Explanation Generation Performance

As shown in Table 2, MVAE outperforms all previous methods across all three datasets, which demonstrates the effectiveness of our proposed model. Inspecting the samples generated by previous methods, we discover that their poor BLEU scores stem mainly from the occasional generation of descriptions without concrete meaning or lack of details, suggesting that their methods lack the ability to capture more specific characteristics of users or items, and corroborating our intuition that noisy embeddings may cause a model to generate unconditional natural language expressions without concrete meaning, since all the explanations are generated by the same decoder but different input embeddings. Moreover, we find that such low-quality predicted explanations usually correspond to users or items with fewer pertinent training samples, demonstrating our assumption that some user or item embeddings remain insufficiently trained.

We further provide a detailed evaluation assessing the quality of explanations for users with different amounts of training samples in Fig. 5. As we can see, our methods improve the quality of explanations with a larger absolute improvement when fewer relevant training samples are present (note the different slope of means of different methods), which suggests that our model can better handle less well trained user and item embeddings. This confirms that our VAE architecture is able to filter out noise and retain meaningful information for the decoder to generate more specific explanations.

4.7 Ablation Study

For an in-depth analysis of the effectiveness of our proposed techniques, as shown in Table 3, we compare our model with two variants introduced earlier. As we can see, the performance of MVAE-NoMLF

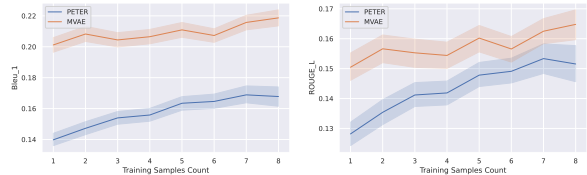


Figure 5: The mean and 95% confidence interval of BLUE-1 and ROUGE-L scores of explanations generated by PETER and MVAE on the *Yelp* dataset. The x-axes represent the count of relevant training samples.

drops substantially. We believe this is because MLF decides how the prior network can be optimized by learning from the recognition network. Also, it controls the fine-grained information that is provided to the reconstruction network. Removing the MLF significantly harms the effectiveness of learning the prior network for fine-grained information. For MVAE-NoKL, with the optimization of representing fine-grained information removed, it is hard for the prior network to model the fine-grained information from the recognition network. Therefore, the model may obtain poor results in testing. In fact, we observe that MVAE-NoKL attains lower training losses in training but has higher testing losses, indicating a significant disparity of distributions between the prior and recognition networks, which degrades the model performance in testing.

4.8 Analysis of MLF

We further examine in detail the necessity and rationality of our proposed MLF. In previous methods, the randomly initialized input embeddings are leveraged by the model directly. However, noisy inputs in the initial training may impede the ability of the model to leverage them and lead to convergence to a sub-optimal solution. We suspect the alternative of simply supplying additional information directly may facilitate the training of the model but result in a large performance gap between training and testing. To confirm our conjecture, we further propose two variants of our model named MVAE-NoRN and MVAE-NoKL. For MVAE-NoRN, we train our model with the testing phase architecture illustrated in Fig. 2(b), i.e., it is trained without the help of ground-truth information directly. For MVAE-NoKL, we replace the $r z_j$ with $\mu_{r z_j}$ to supply fine-grained information to the reconstruction network and replace the TKL with the mean squared error between $\mu_{r z_j}$ and $\mu_{p z_j}$. Under this setting, the additional noise injected into ground-truth information is removed. We compare the re-

	BLEU-1	BLEU-4	ROUGE-1	ROUGE-L	METEOR
MVAE-NoKL	21.03 (↓1.82%)	2.02 (↓10.26%)	18.67 (↓0.55%)	15.15 (↓0.82%)	7.01 (↓3.44%)
MVAE-NoMLF	19.12 (↓10.74%)	1.56 (↓30.70%)	17.95 (↓4.38%)	14.57 (↓4.66%)	6.73 (↓7.30%)
MVAE	21.42	2.25	18.77	15.28	7.26

Table 3: Performance comparison of variants of our model on *Yelp* dataset. Deterioration of the performance is calculated as the relative drop compared with MVAE.

Reference	The <u>staffs</u> are super <u>knowledgeable</u> and obviously care very deeply about the needs and <u>preferences</u> of their <u>customers</u> .
NETE	The service is great.
PETER	The <u>staffs</u> are very friendly and willing to help.
MVAE	The <u>staffs</u> are <u>knowledgeable</u> and the <u>customer service</u> is impressive.
Reference	The <u>atmosphere</u> is <u>relaxing</u> and enjoyable and <u>music</u> made people <u>feel at ease</u> .
NETE	The environment is clear.
PETER	The food is good and the <u>staffs</u> are friendly.
MVAE	The <u>atmosphere</u> and the <u>music</u> are pleasant.

Table 4: Examples of generated explanation by various methods. Fine-grained features are underlined.

sulting training and validation losses in Fig. 6. The training losses of MVAE-NoRN decrease faster in the early stage of optimization, but this soon stagnates and barely improves any further, suggesting that external guided signals are necessary to overcome this plateau, as the prior network without the guidance of the recognition may be unable to distinguish meaningful information from noisy inputs. The MVAE-NoKL model has much lower training losses but higher validation losses, reflecting a large performance gap between training and validation. In contrast, MVAE has reasonable training losses and the lowest validation losses, which implies that the MLF in our model narrows the performance gap between training and validation, proving the effectiveness of our proposed MLF.

4.9 Qualitative Case Study

To further compare the generation quality of explanations generated by previous work and our

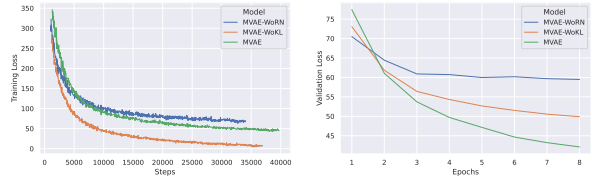


Figure 6: Loss plots: (a) is the training loss and (b) is the validation loss of each model on the *Yelp* dataset.

model, we provide examples in Table 4. We observe that our methods can capture more specific characteristics, thus generating more concrete explanations. For instance, the generated explanation of our model describes fine-grained aspects such as “staff” and “customer service”, which are possible reasons of a recommendation. In contrast, the previous state-of-the-art model PETER only emphasizes the “staff” without a high-level summary on “service”.

5 Conclusion

We present MVAE, a novel model for explanation generation in recommender systems, which has a prior network that eliminates noise while retaining meaningful signals in the input and a recognition network serving as the source of information to guide the learning of the prior network. Further, we propose a Multi-scale distribution Learning Framework along with TKL to prompt this process. Extensive experiments demonstrate the effectiveness of our method and confirm that it can generate high-quality explanations.

Acknowledgements

This work was supported by the National Innovation 2030 Major S&T Project of China (No. 2020AAA0104200 & 2020AAA0104205), National Natural Science Foundation of China (No. 62006077), and Shanghai Sailing Program (No. 20YF1411800).

References

- Alexander A. Alemi, Ian Fischer, Joshua V. Dillon, and Kevin Murphy. 2017. [Deep variational information bottleneck](#). In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.
- Satanjeev Banerjee and Alon Lavie. 2005. [METEOR: An automatic metric for MT evaluation with improved correlation with human judgments](#). In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.
- David Barber and Felix V. Agakov. 2003. [Information maximization in noisy channels : A variational approach](#). In *Advances in Neural Information Processing Systems 16 [Neural Information Processing Systems, NIPS 2003, December 8-13, 2003, Vancouver and Whistler, British Columbia, Canada]*, pages 201–208. MIT Press.
- Ziqiang Cao, Wenjie Li, Sujian Li, and Furu Wei. 2018. [Retrieve, rerank and rewrite: Soft template based neural summarization](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 152–161, Melbourne, Australia. Association for Computational Linguistics.
- Zhongxia Chen, Xiting Wang, Xing Xie, Tong Wu, Guoqing Bu, Yining Wang, and Enhong Chen. 2019. [Co-attentive multi-task learning for explainable recommendation](#). In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, pages 2137–2143. ijcai.org.
- Li Dong, Shaohan Huang, Furu Wei, Mirella Lapata, Ming Zhou, and Ke Xu. 2017. [Learning to generate product reviews from attributes](#). In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 623–632, Valencia, Spain. Association for Computational Linguistics.
- Luciano Floridi and Massimo Chiriatti. 2020. Gpt-3: Its nature, scope, limits, and consequences. *Minds and Machines*, 30(4):681–694.
- Junxian He, Daniel Spokoyny, Graham Neubig, and Taylor Berg-Kirkpatrick. 2019. [Lagging inference networks and posterior collapse in variational autoencoders](#). In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.
- Dan Hendrycks and Kevin Gimpel. 2016. [Gaussian error linear units \(gelus\)](#). *ArXiv preprint*, abs/1606.08415.
- Irina Higgins, Loïc Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. 2017. [beta-vae: Learning basic visual concepts with a constrained variational framework](#). In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.
- Diederik P. Kingma and Jimmy Ba. 2015. [Adam: A method for stochastic optimization](#). In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.
- Diederik P. Kingma and Max Welling. 2014. [Auto-encoding variational bayes](#). In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*.
- Lei Li, Yongfeng Zhang, and Li Chen. 2020. [Generate neural template explanations for recommendation](#). In *CIKM '20: The 29th ACM International Conference on Information and Knowledge Management, Virtual Event, Ireland, October 19-23, 2020*, pages 755–764. ACM.
- Lei Li, Yongfeng Zhang, and Li Chen. 2021. [Personalized transformer for explainable recommendation](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4947–4957, Online. Association for Computational Linguistics.
- Piji Li, Wai Lam, Lidong Bing, and Zihao Wang. 2017a. [Deep recurrent generative decoder for abstractive text summarization](#). In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2091–2100, Copenhagen, Denmark. Association for Computational Linguistics.
- Piji Li, Zihao Wang, Zhaochun Ren, Lidong Bing, and Wai Lam. 2017b. [Neural rating regression with abstractive tips generation for recommendation](#). In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, Shinjuku, Tokyo, Japan, August 7-11, 2017*, pages 345–354. ACM.
- Chin-Yew Lin. 2004. [ROUGE: A package for automatic evaluation of summaries](#). In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. [Bleu: a method for automatic evaluation of machine translation](#). In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

- Iulian Vlad Serban, Alessandro Sordani, Ryan Lowe, Laurent Charlin, Joelle Pineau, Aaron C. Courville, and Yoshua Bengio. 2017. [A hierarchical latent variable encoder-decoder model for generating dialogues](#). In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, pages 3295–3301. AAAI Press.
- Huajie Shao, Shuochao Yao, Dachun Sun, Aston Zhang, Shengzhong Liu, Dongxin Liu, Jun Wang, and Tarek F. Abdelzaher. 2020. [Controlvae: Controllable variational autoencoder](#). In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 8655–8664. PMLR.
- Shikhar Sharma, Layla El Asri, Hannes Schulz, and Jeremie Zumer. 2017. [Relevance of unsupervised metrics in task-oriented dialogue for evaluating natural language generation](#). *ArXiv preprint*, abs/1706.09799.
- Ravid Shwartz-Ziv and Naftali Tishby. 2017. [Opening the black box of deep neural networks via information](#). *ArXiv preprint*, abs/1703.00810.
- Casper Kaae Sønderby, Tapani Raiko, Lars Maaløe, Søren Kaae Sønderby, and Ole Winther. 2016. [Ladder variational autoencoders](#). In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 3738–3746.
- Nava Tintarev and Judith Masthoff. 2015. *Explaining Recommendations: Design and Evaluation*, pages 353–382. Springer US, Boston, MA.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. [Attention is all you need](#). In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.
- Pascal Vincent, H. Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. 2008. Extracting and composing robust features with denoising autoencoders. In *ICML '08*.
- Pascal Vincent, H. Larochelle, Isabelle Lajoie, Yoshua Bengio, and Pierre-Antoine Manzagol. 2010. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *J. Mach. Learn. Res.*, 11:3371–3408.
- Hongwei Wang, Fuzheng Zhang, Jialin Wang, Miao Zhao, Wenjie Li, Xing Xie, and Minyi Guo. 2018a. [Ripplenet: Propagating user preferences on the knowledge graph for recommender systems](#). In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management, CIKM 2018, Torino, Italy, October 22-26, 2018*, pages 417–426. ACM.
- Xiting Wang, Yiru Chen, Jie Yang, Le Wu, Zhengtao Wu, and Xing Xie. 2018b. A reinforcement learning framework for explainable recommendation. In *2018 IEEE international conference on data mining (ICDM)*, pages 587–596. IEEE.
- Tsung-Hsien Wen, Yishu Miao, Phil Blunsom, and Steve J. Young. 2017. [Latent intention dialogue models](#). In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 3732–3741. PMLR.
- Yikun Xian, Zuohui Fu, S. Muthukrishnan, Gerard de Melo, and Yongfeng Zhang. 2019. [Reinforcement knowledge graph reasoning for explainable recommendation](#). In *Proceedings of SIGIR 2019*, pages 285–294, New York, NY, USA. ACM.
- Yuichi Yoshida and Takeru Miyato. 2017. [Spectral norm regularization for improving the generalizability of deep learning](#). *ArXiv preprint*, abs/1705.10941.
- Yongfeng Zhang and Xu Chen. 2020. Explainable recommendation: A survey and new perspectives. In *Foundations and Trends in Information Retrieval*.
- Yongfeng Zhang, Guokun Lai, Min Zhang, Yi Zhang, Yiqun Liu, and Shaoping Ma. 2014. [Explicit factor models for explainable recommendation based on phrase-level sentiment analysis](#). In *The 37th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '14, Gold Coast, QLD, Australia - July 06 - 11, 2014*, pages 83–92. ACM.
- Tiancheng Zhao, Ran Zhao, and Maxine Eskenazi. 2017. [Learning discourse-level diversity for neural dialog models using conditional variational autoencoders](#). In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 654–664, Vancouver, Canada. Association for Computational Linguistics.