# Context-dependent Token-wise Variational Autoencoder for Topic Modeling

Tomonari Masada[1][0000−0002−8358−3699]

Nagasaki University, 1-14 Bunkyo-machi, Nagasaki-shi, Nagasaki, Japan
masada@nagasaki-u.ac.jp

**Abstract.** This paper proposes a new variational autoencoder (VAE) for topic models. The variational inference (VI) for Bayesian models approximates the true posterior distribution by maximizing a lower bound of the log marginal likelihood. We can implement VI as VAE by using a neural network, called encoder, and running it over observations to produce the approximate posterior parameters. However, VAE often suffers from *latent variable collapse*, where the approximate posterior degenerates to the local optima just mimicking the prior due to the over-minimization of the KL-divergence between the approximate posterior and the prior. To address this problem for topic modeling, we propose a new VAE. Since we marginalize out topic probabilities by following the method of Mimno et al., our VAE minimizes a KL-divergence that has not been considered in the existing VAE. Further, we draw samples from the variational posterior for each word token separately. This sampling for Monte-Carlo integration is performed with the Gumbel-softmax trick by using a document-specific context information. We empirically investigated if our new VAE could mitigate the difficulty arising from latent variable collapse. The experimental results showed that our VAE improved the existing VAE for a half of the data sets in terms of perplexity or of normalized pairwise mutual information.

**Keywords:** Topic modeling · Deep learning · Variational autoencoder.

## 1 Introduction

Topic modeling is a well-known text analysis method adopted not only in Web data mining but also in a wide variety of research disciplines, including social sciences[1] and digital humanities.[2] Topic modeling extracts *topics* from a large document set. Topics in topic modeling are defined as categorical distribution over vocabulary words. Each topic is expected to put high probabilities on the words corresponding to some clear-cut semantic contents delivered by the document set. For example, the news articles talking about science may use the words that are rarely used by the articles talking about politics, and vice versa. Topic

[1] https://www.structuraltopicmodel.com/
[2] https://www.gale.com/intl/primary-sources/digital-scholar-lab

modeling can explicate such a difference by providing two different word categorical distributions, one putting high probabilities on the words like experiment, hypothesis, laboratory, etc, and another on the words like government, vote, parliament, etc. Topic modeling achieves two things at the same time. First, it provides as topics categorical distributions defined over words as stated above. Second, topic modeling provides topic proportions for each document. Topic proportions are expected to be effective as a lower-dimensional representation of documents, which reflects the difference of their semantic contents.

This paper discusses the inference for Bayesian topic models. The Bayesian inference aims to infer the posterior distribution. The sampling-based approaches like MCMC infer the posterior by drawing samples from it. While samples are drawn from the true posterior, we can only access the posterior through the drawn samples. This paper focuses on the variational inference (VI), which approximates the posterior with a tractable surrogate distribution. Our task in VI is to estimate the parameters of the surrogate distribution. VI maximizes a lower bound of the log marginal likelihood $\log p(\mathbf{x}_d)$ for every document $\mathbf{x}_d$. A lower bound of $\log p(\mathbf{x}_d)$ is obtained by applying Jensen's inequality as follows:

$$\begin{aligned}\log p(\mathbf{x}_d) &= \log \int p(\mathbf{x}_d|\mathbf{\Phi})p(\mathbf{\Phi})d\mathbf{\Phi} = \log \int q(\mathbf{\Phi})\frac{p(\mathbf{x}_d|\mathbf{\Phi})p(\mathbf{\Phi})}{q(\mathbf{\Phi})}d\mathbf{\Phi} \\ &\geq \int q(\mathbf{\Phi}) \log \frac{p(\mathbf{x}_d|\mathbf{\Phi})p(\mathbf{\Phi})}{q(\mathbf{\Phi})}d\mathbf{\Phi} = \mathbb{E}_{q(\mathbf{\Phi})}\big[\log p(\mathbf{x}_d|\mathbf{\Phi})\big] - \mathrm{KL}(q(\mathbf{\Phi}) \parallel p(\mathbf{\Phi})) \equiv \mathcal{L}_d\end{aligned} \tag{1}$$

where $\mathbf{\Phi}$ are the model parameters and $q(\mathbf{\Phi})$ a surrogate distribution approximating the true posterior. This lower bound is often called ELBO (Evidence Lower BOund), which we denote by $\mathcal{L}_d$.

In this paper, we consider the variational autoencoder (VAE), a variational inference using neural networks for approximating the true posterior. To estimate the parameters of the approximate posterior $q(\mathbf{\Phi})$ in Eq. (1), we can use a neural network called *encoder* as follows. By feeding each observed document $\mathbf{x}_d$ as an input to the encoder, we obtain the parameters of the document-dependent approximate posterior $q(\mathbf{\Phi}|\mathbf{x}_d)$ as the corresponding output. The weights and biases of the encoder network is shared by all documents. Further, we approximate the ELBO in Eq. (1) by using Monte Carlo integration:

$$\mathcal{L}_d \approx \frac{1}{S}\sum_{s=1}^{S} \log p(\mathbf{x}_d|\hat{\mathbf{\Phi}}^{(s)}) - \frac{1}{S}\sum_{s=1}^{S} \log \frac{q(\hat{\mathbf{\Phi}}^{(s)}|\mathbf{x}_d)}{p(\hat{\mathbf{\Phi}}^{(s)})} \tag{2}$$

where $\hat{\mathbf{\Phi}}^{(s)}$ are samples drawn from $q(\mathbf{\Phi}|\mathbf{x}_d)$. The sampling $\hat{\mathbf{\Phi}} \sim q(\mathbf{\Phi}|\mathbf{x}_d)$ can be performed by using reparameterization trick [10, 18, 20, 24]. For example, if we choose approximate posteriors from among the multivariate diagonal Gaussian distributions $\mathcal{N}(\boldsymbol{\mu}(\mathbf{x}_d), \boldsymbol{\sigma}(\mathbf{x}_d))$ of dimension $K$, we use an encoder with $2K$ outputs, the one half of which are used as mean parameters $\boldsymbol{\mu}(\cdot)$ and the other half as log standard deviation parameters $\log \boldsymbol{\sigma}(\cdot)$. We then obtain a sample $\hat{\mathbf{\Phi}} \sim \mathcal{N}(\boldsymbol{\mu}(\mathbf{x}_d), \boldsymbol{\sigma}(\mathbf{x}_d))$ by using a sample $\hat{\boldsymbol{\epsilon}}$ from the $K$-dimensional standard

Gaussian distribution as $\hat{\mathbf{\Phi}} = \hat{\boldsymbol{\epsilon}} \odot \boldsymbol{\sigma}(\mathbf{x}_d) + \boldsymbol{\mu}(\mathbf{x}_d)$, where $\odot$ is the element-wise multiplication. The reparameterization trick enables us to backpropagate to the weights and biases defining the encoder network through samples. While $\mathcal{L}_d$ in Eq. (2) is formulated for a particular document $\mathbf{x}_d$, we maximize the ELBO for many documents simultaneously by using mini-batch optimization.

However, VAE often suffers from *latent variable collapse*, where the approximate posterior $q(\mathbf{\Phi}|\mathbf{x}_d)$ degenerates to the local optima just mimicking the prior $p(\mathbf{\Phi})$. When the maximization of $\mathcal{L}_d$ in Eq. (1) makes the KL-divergence term $\mathrm{KL}(q(\mathbf{\Phi}) \parallel p(\mathbf{\Phi}))$ excessively close to zero, latent variable collapse occurs. The main contribution of this paper is to provide a new VAE for topic modeling, where we minimize a KL-divergence that has not been considered in the previous VAE. While we only consider latent Dirichlet allocation (LDA) [1], a similar proposal can be given also for other topic models. The evaluation experiment, conducted over four large document sets, showed that the proposed VAE improved the previous VAE for some data sets in terms of perplexity or of normalized pairwise mutual information (NPMI).

The rest of the paper is organized as follows. Section 2 introduces LDA and describes the details of our proposal. Section 3 gives the results of the experiment where we compare our new VAE with other variational inference methods. Section 4 provides relevant previous work. Section 5 concludes the paper by suggesting future research directions.

## 2 Method

### 2.1 Latent Dirichlet Allocation (LDA)

This subsection introduces LDA [1], for which we propose a new VAE. Assume that there are $D$ documents $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_D\}$. We denote the vocabulary size by $V$ and the vocabulary set by $\{w_1, \ldots, w_V\}$. Each document $\mathbf{x}_d$ is represented as bag-of-words, i.e., a multiset of vocabulary words. The total number of the word tokens in $\mathbf{x}_d$ is referred to by $n_d$. Let $x_{d,i}$ be the observable random variable giving the word appearing as the $i$-th token in $\mathbf{x}_d$. That is, $x_{d,i} = w_v$ means that a token of the word $w_v$ appears as the $i$-th token in $\mathbf{x}_d$. Topic modeling assigns each word token to one among the fixed number of topics. This can be regarded as a token-level clustering. We denote the number of topics by $K$ and the topic set by $\{t_1, \ldots, t_K\}$. Let $z_{d,i}$ be the latent random variable representing the topic to which $x_{d,i}$ is assigned. That is, $z_{d,i} = t_k$ means that the $i$-th word token in $\mathbf{x}_d$ is assigned to the topic $t_k$.

LDA is a generative model, which generates the data set $\mathbf{X}$ as below. First of all, we assume that a categorical distribution $\mathrm{Categorical}(\boldsymbol{\beta}_k)$ is prepared for each topic $t_k$, where $\beta_{k,v}$ is the probability of the word $w_v$ in the topic $t_k$. $\boldsymbol{\beta}_k = (\beta_{k,1}, \ldots, \beta_{k,V})$ satisfies $\sum_v \beta_{k,v} = 1$. LDA generates documents as follows:

- For $d = 1, \ldots, D$:
  - A parameter vector $\boldsymbol{\theta}_d = (\theta_{d,1}, \ldots, \theta_{d,K})$ of the categorical distribution $\mathrm{Categorical}(\boldsymbol{\theta}_d)$ is drawn from the symmetric Dirichlet prior $\mathrm{Dirichlet}(\alpha)$.

- For $i = 1, \ldots, n_d$:
  - A latent topic $z_{d,i}$ is drawn from Categorical($\boldsymbol{\theta}_d$)
  - A word $x_{d,i}$ is drawn from Categorical($\boldsymbol{\beta}_{z_{d,i}}$).

where $\theta_{d,k}$ is the probability of $t_k$ in $\mathbf{x}_d$. In this paper, we apply no prior distribution to the categorical distributions Categorical($\boldsymbol{\beta}_k$) for $k = 1, \ldots, K$.

Based on the above generative story, the marginal likelihood (also termed the evidence) of each document $\mathbf{x}_d$ for $d = 1, \ldots, D$ can be given as follows:

$$p(\mathbf{x}_d; \alpha, \boldsymbol{\beta}) = \int \Big( \sum_{\mathbf{z}_d} p(\mathbf{x}_d|\mathbf{z}_d; \boldsymbol{\beta}) p(\mathbf{z}_d|\boldsymbol{\theta}_d) \Big) p(\boldsymbol{\theta}_d; \alpha) d\boldsymbol{\theta}_d \tag{3}$$

where $p(\boldsymbol{\theta}_d; \alpha)$ represents the prior Dirichlet($\alpha$). The variational inference (VI) given in [1] maximizes the following lower bound of the log evidence:

$$\begin{aligned} \log p(\mathbf{x}_d; \alpha, \boldsymbol{\beta}) \geq \mathbb{E}_{q(\mathbf{z}_d)} \big[ \log p(\mathbf{x}_d|\mathbf{z}_d; \boldsymbol{\beta}) \big] + \mathbb{E}_{q(\boldsymbol{\theta}_d) q(\mathbf{z}_d)} \big[ \log p(\mathbf{z}_d|\boldsymbol{\theta}_d) \big] \\ - \mathrm{KL}(q(\boldsymbol{\theta}_d) \parallel p(\boldsymbol{\theta}_d; \alpha)) - \mathbb{E}_{q(\mathbf{z}_d)} \big[ \log q(\mathbf{z}_d) \big] \end{aligned} \tag{4}$$

This lower bound is obtained by applying Jensen's inequality as in Eq. (1) and is often called ELBO (Evidence Lower BOund).

In the VI given by [1], the variational distribution $q(\boldsymbol{\theta}_d)$ is a Dirichlet distribution parameterized separately for each document. That is, there is no connection between the variational parameters of different documents. By replacing this per-document variational posterior with an amortized variational posterior $q(\boldsymbol{\theta}_d|\mathbf{x}_d)$, we obtain a variational autoencoder (VAE) for LDA [14, 22]. However, VAE often leads to a problem called *latent variable collapse*, where the KL-divergence term in Eq.(4) is collapsed toward zero. Latent variable collapse makes the approximate posterior $q(\boldsymbol{\theta}_d|\mathbf{x}_d)$ almost equivalent to the prior $p(\boldsymbol{\theta}_d; \alpha)$. Consequently, we cannot obtain informative topic probabilities $\boldsymbol{\theta}_d$ for any document.

### 2.2 Context-dependent and Token-wise VAE for LDA

The VAE for LDA proposed in [22] avoids latent variable collapse only by tweaking the optimization, i.e., by tuning optimizer parameters, clipping the gradient, adopting batch normalization, and so on. There are also other proposals using KL-cost annealing [8, 21] to avoid latent variable collapse. In contrast, we address the issue by expelling the problematic KL-divergence term $\mathrm{KL}(q(\boldsymbol{\theta}_d) \parallel p(\boldsymbol{\theta}_d; \alpha))$ from ELBO. We follow Mimno et al. [15] and marginalize out the per-document topic proportions $\boldsymbol{\theta}_d$ in Eq. (3) and obtain the following evidence:

$$p(\mathbf{x}_d; \alpha, \boldsymbol{\beta}) = \sum_{\mathbf{z}_d} p(\mathbf{x}_d|\mathbf{z}_d; \boldsymbol{\beta}) p(\mathbf{z}_d; \alpha) \tag{5}$$

By marginalizing out $\boldsymbol{\theta}_d$, we can avoid the approximation of the posterior distribution over $\boldsymbol{\theta}_d$. Consequently, we don't need to estimate the problematic KL-divergence. However, we now need to approximate the posterior over the topic

assignments $\mathbf{z}_d$. The ELBO to be maximized in our case is

$$\begin{aligned}\log p(\mathbf{x}_d|\alpha,\boldsymbol{\beta}) &= \log \sum_{\mathbf{z}_d} p(\mathbf{x}_d|\boldsymbol{\beta},\mathbf{z}_d)p(\mathbf{z}_d|\alpha) \\ &\geq \mathbb{E}_{q(\mathbf{z}_d|\mathbf{x}_d)}\big[\log p(\mathbf{x}_d|\boldsymbol{\beta},\mathbf{z}_d)\big] - \mathrm{KL}(q(\mathbf{z}_d|\mathbf{x}_d) \parallel p(\mathbf{z}_d|\alpha)) \end{aligned} \tag{6}$$

The KL-divergence $\mathrm{KL}(q(\mathbf{z}_d|\mathbf{x}_d) \parallel p(\mathbf{z}_d|\alpha))$ our VAE minimizes has not been considered in the previous VAE for topic modeling. We found in our evaluation experiment that this KL-divergence required no special treatment for achieving a good evaluation result. This finding is the main contribution.

The first part of the left hand side of Eq. (6) can be rewritten as

$$\mathbb{E}_{q(\mathbf{z}_d|\mathbf{x}_d)}\big[\log p(\mathbf{x}_d|\boldsymbol{\beta},\mathbf{z}_d)\big] = \sum_{i=1}^{N_d} \mathbb{E}_{q(z_{d,i}|\mathbf{x}_d)}\big[\log \beta_{z_{d,i},x_{d,i}}\big] \tag{7}$$

where we assume that $q(\mathbf{z}_d|\mathbf{x}_d)$ is factorized as $\prod_{i=1}^{n_d} q(z_{d,i}|\mathbf{x}_d)$ and regard each component $q(z_{d,i}|\mathbf{x}_d)$ as categorical distribution. In this paper, we parameterize $\beta_{k,v}$ as $\beta_{k,v} \propto \exp(\eta_{k,v} + \eta_{0,v})$. The variable $\eta_{0,v}$ is shared by all topics and roughly represents the background word probabilities. The per-topic parameters $\eta_{k,v}$ and the topic-indifferent parameters $\eta_{0,v}$ are estimated directly by maximizing the likelihood in Eq. (7).

The negative KL-divergence in Eq. (6) is given as $-\mathrm{KL}(q(\mathbf{z}_d|\mathbf{x}_d) \parallel p(\mathbf{z}_d|\alpha)) = \mathbb{E}_{q(\mathbf{z}_d|\mathbf{x}_d)}\big[\log p(\mathbf{z}_d|\alpha)\big] - \mathbb{E}_{q(\mathbf{z}_d|\mathbf{x}_d)}\big[\log q(\mathbf{z}_d|\mathbf{x}_d)\big]$. The first half is given as follows by noticing that $p(\mathbf{z}_d|\alpha)$ is Pólya distribution:

$$\mathbb{E}_{q(\mathbf{z}_d|\mathbf{x}_d)}\big[\ln p(\mathbf{z}_d|\alpha)\big] = \mathbb{E}_{q(\mathbf{z}_d|\mathbf{x}_d)}\left[\ln\left\{\frac{(N_d!)\Gamma(K\alpha)}{\Gamma(N_d+K\alpha)}\prod_{k=1}^{K}\frac{\Gamma(n_{d,k}+\alpha)}{(n_{d,k}!)\Gamma(\alpha)}\right\}\right] \tag{8}$$

where $n_{d,k}$ is the number of the word tokens assigned to the topic $t_k$ in $\mathbf{x}_d$. It should be noted that $n_{d,k}$ depends on $\mathbf{z}_d$. The second half of the KL-divergence is an entropy term:

$$-\mathbb{E}_{q(\mathbf{z}_d|\mathbf{x}_d)}\big[\ln q(\mathbf{z}_d|\mathbf{x}_d)\big] = -\sum_{i=1}^{N_d} \mathbb{E}_{q(z_{d,i}|\mathbf{x}_d)}\big[\log q(z_{d,i}|\mathbf{x}_d)\big] \tag{9}$$

The main task in our proposed VAE is to estimate the above three expectations by drawing samples from the approximate posterior $q(\mathbf{z}_d|\mathbf{x}_d)$.

Our proposed method models the approximate posterior $q(\mathbf{z}_d|\mathbf{x}_d)$ in an amortized manner by using multilayer perceptron (MLP). As already stated, we assume that $q(\mathbf{z}_d|\mathbf{x}_d)$ is factorized as $\prod_{i=1}^{n_d} q(z_{d,i}|\mathbf{x}_d)$ and regard each component $q(z_{d,i}|\mathbf{x}_d)$ as categorical distribution. The parameters of each $q(z_{d,i}|\mathbf{x}_d)$ for $i = 1, \ldots, n_d$ are obtained as follows by using an encoder MLP. Let $\mathbf{e}_v$ denote the embedding vector of the vocabulary word $w_v$. Let $\boldsymbol{\gamma}_{d,i} = (\gamma_{d,i,1}, \ldots, \gamma_{d,i,K})$ denote the parameter vector of the categorical distribution $q(z_{d,i}|\mathbf{x}_d)$. $\gamma_{d,i,k}$ is the probability that the $i$-th token in the document $\mathbf{x}_d$ is assigned to the topic $t_k$. We

**Table 1.** Specifications of data sets

| | $D_{train}$ | $D_{valid}$ | $D_{test}$ | $V$ |
|---|---|---|---|---|
| DBLP | 2,779,431 | 346,792 | 347,291 | 155,187 |
| PUBMED | 5,738,672 | 1,230,981 | 1,230,347 | 141,043 |
| DIET | 10,549,675 | 1,318,663 | 1,318,805 | 152,200 |
| NYTimes | 239,785 | 29,978 | 29,968 | 102,660 |

then obtain $\boldsymbol{\gamma}_{d,i}$ as the softmaxed output of the encoder MLP, whose input is the concatenation of the embedding vector of the $i$-th word token, i.e., $\mathbf{e}_{x_{d,i}}$, and the mean of the embedding vectors of all word tokens in $\mathbf{x}_d$, i.e., $\bar{\mathbf{e}}_d \equiv \frac{1}{n_d}\sum_{i=1}^{n_d}\mathbf{e}_{x_{d,i}}$. That is, $\boldsymbol{\gamma}_{d,i} = \mathrm{Softmax}(\mathrm{MLP}(\mathbf{e}_{x_{d,i}}, \bar{\mathbf{e}}_d))$. To make the posterior inference memory efficient, we reuse $\eta_{k,v}$, which defines $\beta_{k,v}$, as $e_{v,k}$, i.e., the $k$-th element of the embedding vector of the word $w_v$. Such weight tying is also used in [16, 17].

We then draw samples from $q(z_{d,i}|\mathbf{x}_d) \equiv \mathrm{Categorical}(\boldsymbol{\gamma}_{d,i})$ with the Gumbel-softmax trick [7]. That is, a discrete sample $z_{d,i} \equiv \mathrm{argmax}_k\left(g_{d,i,k} + \log\gamma_{d,i,k}\right)$ is approximated by the continuous sample vector $\mathbf{y}_{d,i} = (y_{d,i,1}, \ldots, y_{d,1,K})$, i.e., $y_{d,i,k} \propto \exp((g_{d,i,k} + \log\gamma_{d,i,k})/\tau)$ for $k = 1, \ldots, K$, where $g_{d,i,1}, \ldots, g_{d,i,K}$ are the samples from the standard Gumbel distribution $\mathrm{Gumbel}(0, 1)$ and $\tau$ the temperature. We can obtain $n_{d,k}$ in Eq. (8) as $n_{d,k} = \sum_{i=1}^{n_d} y_{d,i,k}$. It should be noted that we obtain a different sample for each token. That is, the sampling from the approximate posterior is performed in a token-wise manner. Further, recall that our encoder MLP accepts context information, i.e., $\bar{\mathbf{e}}_d$. Therefore, we call our method *context-dependent token-wise* variational autoencoder (CTVAE). While we here propose CTVAE only for LDA, a similar proposal can be made for other Bayesian topic models. This would be interesting as future work.

## 3 Evaluation Experiment

### 3.1 Data Sets and Compared Methods

The evaluation used three large document sets to demonstrate both effectiveness and scalability. The specifications are given in Table 1, where $D_{train}$, $D_{valid}$, and $D_{test}$ are the sizes of training set, validation set, and test set, respectively, and $V$ the vocabulary size. DBLP is a subset of the records downloaded from DBLP Web site[3] on July 24, 2018, where each paper title is regarded as document. PUBMED is a part of the bag-of-words data set provided by the UC Irvine Machine Learning Repository.[4] DIET data set is a subset of the Diet Record of both houses of the National Diet of Japan from 1993 to 2012,[5] where each statement by the speaker is regarded as a separate document. Since DIET data set is given in Japanese, we used MeCab morphological analyzer.[6] We also used the NYTimes part from the above-mentioned beg-of-words data set.

[3] https://dblp.uni-trier.de/xml/dblp.xml.gz
[4] https://archive.ics.uci.edu/ml/datasets/bag+of+words
[5] http://kokkai.ndl.go.jp/
[6] http://taku910.github.io/mecab/

Our proposal, CTVAE, and all compared methods were implemented in PyTorch.[7] The number of topics was fixed to $K = 50$. All free parameters, including the number of layers and the layer sizes of the encoder MLP, were tuned based on the perplexity computed over the validation set. The perplexity is defined as the exponential of the negative log likelihood of each document, where the log likelihood is normalized by the number of word tokens. The mean of the perplexity over all validation documents was used as the evaluation measure for tuning free parameters. We adopted Adam optimizer [9], where the learning rate scheduling was also tuned based on the validation perplexity.

We compared our proposal to the following three variational inference methods: a plain variational autoencoder (VAE), mini-batch variational Bayes (VB), and adversarial variational Bayes (AVB). Each is explained below.

**VAE.** We proposed CTVAE in order to avoid latent variable collapse in the existing VAE. Therefore, we compared our proposal to a plain implementation of VAE. While a version of VAE for topic modeling has already been proposed [22], it involves additional elaborations that are not relevant to our comparison. Therefore, we implemented VAE in a more elementary manner by following the neural topic model described in [13]. We adopted the standard multivariate Gaussian distribution as the prior distribution. The topic proportions $\boldsymbol{\theta}_d$ were obtained as $\boldsymbol{\theta}_d \equiv \mathrm{Softmax}(\hat{\boldsymbol{\epsilon}} \odot \boldsymbol{\sigma}(\mathbf{x}_d) + \boldsymbol{\mu}(\mathbf{x}_d))$, where $\hat{\boldsymbol{\epsilon}} \sim \mathcal{N}(\mathbf{0}, \mathbf{1})$. When maximizing ELBO, we minimized the KL-divergence from $\mathcal{N}(\boldsymbol{\mu}(\mathbf{x}_d), \boldsymbol{\sigma}(\mathbf{x}_d))$ to $\mathcal{N}(\mathbf{0}, \mathbf{1})$. The mini-batch optimization with Adam was conducted for training the encoder MLP, whose output was the concatenation of $\boldsymbol{\mu}(\mathbf{x}_d)$ and $\log \boldsymbol{\sigma}(\mathbf{x}_d)$, and also for updating other parameters. While the VAE proposed by [22] avoids latent variable collapse only by tweaking the optimization, the VAE in our comparison experiment used KL-cost annealing [21, 8] or path derivative ELBO gradient [19] along with the learning rate scheduling tuned based on the validation perplexity.

**VB.** We also compared our proposal to the variational Bayes (VB) for LDA [1], where no neural network was used. We applied no prior distribution to the per-topic word categorical distributions and directly estimated the per-topic word probabilities by maximizing the data likelihood. We ran a mini-batch version of VB for estimating the per-topic word probabilities. The per-document variational parameters were updated for each document separately in the same manner as [1, 6]. While our implementation was based on [6], the optimization was performed in a 'deep learning' manner. That is, we used Adam optimizer, whose learning rate scheduling was tuned based on the validation perplexity.

**AVB.** We additionaly compared CTVAE to adversarial variational Bayes (AVB) [12]. AVB is a variational inference where we use implicit distribution for approximating the true posterior. Since the approximate posterior is implicit, we cannot obtain an explicit formula of its density function and thus cannot estimate the KL-divergence term in Eq. (1) by the Monte Carlo integration. The special feature of AVB is that a GAN-like discriminative density ratio estimation [4] is adopted for the estimation of the density ratio between the prior and the variational posterior, which is required for the estimation of the KL-divergence.

[7] https://pytorch.org/

**Table 2.** Evaluation results in terms of test set perplexity

| | CTVAE | VAE | AVB | VB |
|---|---|---|---|---|
| DBLP | 1583.88 | 1749.64 | 1326.36 | **1249.92** |
| PUBMED | 3669.79 | 1072.63 | **991.43** | 2958.56 |
| DIET | 309.04 | 364.94 | 336.33 | **219.97** |
| NYTimes | 4077.60 | 2893.23 | **2375.12** | 3155.75 |

Since the AVB described in [12] is not specialized to topic modeling, the AVB proposed for topic modeling [11] was used in our comparison.

### 3.2 Results

Table 2 includes the perplexity computed over the test set for each compared method and each data set. A low perplexity indicates that the approximate posterior is good at predicting the word occurrences in the test documents. CTVAE improved the test set perplexity of VAE both for DBLP and DIET data sets. Especially, for DIET data set, the perplexity of CTVAE was the second best among the four compared methods. It can be said that when VAE does not give any good result, we can consider to use CTVAE as a possible alternative. AVB gave the best perplexity for PUBMED and NYTimes data sets and was always better than VAE. However, AVB has two MLPs, i.e., the encoder and the discriminator, and we need to tune the free parameters of these two MLPs. Therefore, it is more difficult to train AVB than VAE and CTVAE, because both VAE and CTVAE have only one MLP, i.e., the encoder. VB gave the best perplexity for DBLP and DIET data sets. However, VB gave the second worst perplexity for the other two data sets. Further, VB is not efficient in predicting topic probabilities for unseen documents, because this prediction requires tens of iterations for updating approximate posterior parameters until convergence. In contrast, VAE, CTVAE, and AVB uses the encoder MLP for predicting topic probabilities for unseen documents. Therefore, these methods only require a single forward computation over the encoder for this prediction. As for the computational efficiency in time, in case of DIET data set, for example, it took 14 hours for CTVAE, 10 hours for AVB, 7 hours for VAE, and 3 hours for VB to achieve the minimum validation perplexity, respectively. While VAE and AVB only uses a single sample for Monte Carlo estimation of the expectations with respect to the approximate posterior, CTVAE uses 10 samples. Consequently, CTVAE took more hours to train the encoder MLP than VAE and AVB. VB was the most efficient in time, because it used no neural networks.

We next present the results of the evaluation of the high probability words in the extracted topics in Table 3. We measured the quality of the high probability words in each topic with the normalized point-wise mutual information (NPMI) [2]. NPMI is defined for a pair of words $(w, w')$ as $\mathrm{NPMI}(w, w') = \left(\log \frac{p(w,w')}{p(w)p(w')}\right) / -\log p(w, w')$. We calculated the probability $p(w)$ as the number of the test documents containing the word $w$ divided by the total number

**Table 3.** Evaluation results in terms of test set NPMI

| | CTVAE | VAE | AVB | VB |
|---|---|---|---|---|
| DBLP | 0.003 | 0.036 | **0.062** | 0.018 |
| PUBMED | 0.127 | 0.123 | 0.190 | **0.210** |
| DIET | 0.069 | 0.096 | 0.079 | **0.116** |
| NYTimes | 0.141 | 0.105 | 0.075 | **0.218** |

of the test documents and $p(w, w')$ as the number of the test documents containing both $w$ and $w'$ divided by the total number of the test documents. Table 3 includes NPMI computed over the test set by using the top 20 highest probability words from each of the 50 topics. We obtain NPMI for each of the $20(20-1)/2 = 190$ word pairs made by pairing two from the top 20 words and then calculate the average over all word pairs from all 50 topics. NPMI evaluates how well the co-occurrence of high probability words in each extracted topic reflects the actual word co-occurrence in the test documents. While CTVAE was worse than VAE in terms of test perplexity for PUBMED and NYTimes data sets, CTVAE was better in terms of NPMI for these two data sets. Especially, for NYTimes data set, CTVAE gave the second best NPMI. Table 2 showed that AVB was always better than VAE in terms of perplexity. However, AVB improved the NPMI of VAE only for DBLP and PUBMED data sets. Table 3 shows that VB may be the best choice if we are only interested in the quality of the high probability words of extracted topics and are not interested in the prediction capacity. Finally, Table 4 provides an example of high probability words obtained by the compared methods for DBLP data set. We selected five topics from among the 50 topics and presented the top 10 highest probability words for each of the five topics.

## 4 Previous Work

While VAE is widely used in the deep learning field, it is still difficult to apply it to topic modeling, because topic models have discrete latent variables $z_{d,i}$ representing topic assignments. The reparameterization trick [10, 18, 20, 24] cannot be straightforwardly applied to discrete variables. Therefore, low-variance gradient estimation for discrete variables is realized by using some elaborated technique [25] or the Gumbel-softmax trick [7]. The existing proposals of variational inference for topic modeling in the deep learning field [3, 13, 22, 26] marginalize out the discrete variables $z_{d,i}$ and only consider continuous variables. In contrast, CTVAE marginalizes out the topic probabilities. This marginalizing out is also performed in the Gibbs sampling [5], the collapsed variational Bayes [23], and the variational inference proposed by [15]. This paper follows [15] and shows that we can marginalize out the topic probabilities and then propose a new VAE for topic modeling. Since CTVAE marginalizes out the topic probabilities, we don't need to estimate $\mathrm{KL}(q(\boldsymbol{\theta}_d) \parallel p(\boldsymbol{\theta}_d; \alpha))$ in Eq. (4) and thus can avoid the latent variable collapse the existing VAE [14, 22] have suffered from. The VAE proposed

**Table 4.** Top 10 highest-probability words in randomly chosen five topics (DBLP)

| | |
|---|---|
| CTVAE | structure probabilistic game hoc partial bounds project weighted dependent structural |
| | modeling neural energy images object driven processes group code perspective |
| | frequency noise collaborative joint hierarchical open behavior review transmission surface |
| | robots inter convolutional policies symbolic grids stabilization sliding transient play |
| | hybrid graph mining rate features type equations functional some challenges |
| VAE | graphs equations polynomials equation sets matrices finite numbers number polynomial |
| | real time planning software tool carlo models scheduling monte making |
| | encryption secure privacy key authentication authenticated xml queries query efficient |
| | optimization process bayesian prediction evolutionary making time processes using portfolio |
| | codes channels coding fading decoding modulation mimo video channel ofdm |
| VB | network based function radial multi using systems dynamic management neural |
| | order delay automata cellular load finite stability quantum higher first |
| | web random services secure privacy games e key de public |
| | controller fpga design generation based next connectivity driven network intelligent |
| | robot artificial time optimal joint real systematic synthesis filter humanoid |
| AVB | management ad distributed routing editorial wireless hoc network sensor agent |
| | equations nonlinear finite order linear graphs differential equation estimation method |
| | theory finite linear trees minimum reviews codes structure simulation methods |
| | image algorithm fast segmentation detection images coding feature motion video |
| | hoc ad wireless networks radio sensor channel routing energy access |

for topic modeling by [22] only tackles latent variable collapse by tweaking the optimization. In this paper, we proposed a possible way to mitigate the estimation of the problematic KL-divergence. It is also an important contribution of this paper to prove empirically that the Gumbel-softmax trick works for the discrete latent variables $z_{d,i}$ in topic modeling.

## 5 Conclusions

We proposed a new variational autoencoder for topic modeling, called context-dependent token-wise variational autoencoder (CTVAE), by marginalizing out the per-document topic probabilities. The main feature of CTVAE was that the sampling from the variational posterior was performed in a token-wise manner with respect to the context specified by each document. The context was represented as the mean of the embedding vectors of the words each document contains. This feature led to the test perplexity better than the existing VAE for some data sets. However, the improvement was not that remarkable. It is an important future research direction to improve the performance of CTVAE. The experimental results may indicate that we have not yet find a effective way to train the encoder MLP of CTVAE.

## References

1. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet allocation. J. Mach. Learn. Res. **3**, 993–1022 (2003)

2. Bouma, G.: Normalized (Pointwise) Mutual Information in Collocation Extraction. In: From Form to Meaning: Processing Texts Automatically, Proceedings of the Biennial GSCL Conference 2009. pp. 31–40 (2009)
3. Dieng, A.B., Wang, C., Gao, J., Paisley, J.W.: TopicRNN: A recurrent neural network with long-range semantic dependency. CoRR **abs/1611.01702** (2016), http://arxiv.org/abs/1611.01702
4. Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A.C., Bengio, Y.: Generative adversarial nets. In: Advances in Neural Information Processing Systems 27 (NIPS). pp. 2672–2680 (2014)
5. Griffiths, T.L., Steyvers, M.: Finding scientific topics. Proceedings of the National Academy of Sciences **101**(suppl 1), 5228–5235 (2004)
6. Hoffman, M.D., Blei, D.M., Bach, F.R.: Online learning for latent Dirichlet allocation. In: Advances in Neural Information Processing Systems 23 (NIPS). pp. 856–864 (2010)
7. Jang, E., Gu, S., Poole, B.: Categorical reparameterization with Gumbel-softmax. CoRR **abs/1611.01144** (2016), http://arxiv.org/abs/1611.01144
8. Kim, Y., Wiseman, S., Miller, A.C., Sontag, D., Rush, A.M.: Semi-amortized variational autoencoders. In: Proceedings of the 35th International Conference on Machine Learning (ICML). pp. 2683–2692 (2018)
9. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. CoRR **abs/1412.6980** (2014), http://arxiv.org/abs/1412.6980
10. Kingma, D.P., Welling, M.: Auto-encoding variational Bayes. CoRR **abs/1312.6114** (2013), http://arxiv.org/abs/1312.6114
11. Masada, T., Takasu, A.: Adversarial learning for topic models. In: Proceedings of the 14th International Conference on Advanced Data Mining and Applications (ADMA). pp. 292–302 (2018)
12. Mescheder, L.M., Nowozin, S., Geiger, A.: Adversarial variational Bayes: Unifying variational autoencoders and generative adversarial networks. In: Proceedings of the 34th International Conference on Machine Learning (ICML). pp. 2391–2400 (2017)
13. Miao, Y., Grefenstette, E., Blunsom, P.: Discovering discrete latent topics with neural variational inference. In: Proceedings of the 34th International Conference on Machine Learning (ICML). pp. 2410–2419 (2017)
14. Miao, Y., Yu, L., Blunsom, P.: Neural variational inference for text processing. In: Proceedings of the 33rd International Conference on Machine Learning (ICML). pp. 1727–1736 (2016)
15. Mimno, D., Hoffman, M.D., Blei, D.M.: Sparse stochastic inference for latent Dirichlet allocation. In: Proceedings of the 29th International Conference on Machine Learning (ICML). pp. 1515–1522 (2012)
16. Mnih, A., Hinton, G.E.: A scalable hierarchical distributed language model. In: Advances in Neural Information Processing Systems 21 (NIPS). pp. 1081–1088 (2008)
17. Press, O., Wolf, L.: Using the output embedding to improve language models. CoRR **abs/1608.05859** (2016), http://arxiv.org/abs/1608.05859
18. Rezende, D.J., Mohamed, S., Wierstra, D.: Stochastic backpropagation and approximate inference in deep generative models. In: Proceedings of the 31st International Conference on Machine Learning (ICML). pp. II–1278–II–1286 (2014)
19. Roeder, G., Wu, Y., Duvenaud, D.K.: Sticking the landing: Simple, lower-variance gradient estimators for variational inference. In: Advances in Neural Information Processing Systems 30 (NIPS). pp. 6928–6937 (2017)

20. Salimans, T., Knowles, D.A.: Fixed-form variational posterior approximation through stochastic linear regression. CoRR **abs/1206.6679** (2012), http://arxiv.org/abs/1206.6679
21. Sønderby, C.K., Raiko, T., Maaløe, L., Sønderby, S.K., Winther, O.: Ladder variational autoencoders. In: Advances in Neural Information Processing Systems 29 (NIPS). pp. 3738–3746 (2016)
22. Srivastava, A., Sutton, C.: Autoencoding variational inference for topic models. CoRR **abs/1703.01488** (2017), http://arxiv.org/abs/1703.01488
23. Teh, Y.W., Newman, D., Welling, M.: A collapsed variational Bayesian inference algorithm for latent Dirichlet allocation. In: Advances in Neural Information Processing Systems 19 (NIPS). pp. 1353–1360 (2006)
24. Titsias, M.K., Lázaro-Gredilla, M.: Doubly stochastic variational Bayes for non-conjugate inference. In: Proceedings of the 31st International Conference on Machine Learning (ICML). pp. II–1971–II–1980 (2014)
25. Tokui, S., Sato, I.: Reparameterization trick for discrete variables. CoRR **abs/1611.01239** (2016), http://arxiv.org/abs/1611.01239
26. Wang, W., Gan, Z., Wang, W., Shen, D., Huang, J., Ping, W., Satheesh, S., Carin, L.: Topic compositional neural language model. In: Proceedings of the 21st International Conference on Artificial Intelligence and Statistics (AISTATS). pp. 356–365 (2018)