Alleviating "Posterior Collapse" in Deep Topic Models via Policy Gradient

Anonymous Author(s)
Affiliation
Address
email

Abstract

Deep topic models have been proven as a promising way to extract hierarchical latent representations from documents represented as high-dimensional bag-of-words vectors. However, the representation capability of existing deep topic models is still limited by the phenomenon of "*posterior collapse*", which has been widely criticized in deep generative models, resulting in the higher-level latent representations exhibiting similar or meaningless patterns. To this end, in this paper, we first develop a novel deep-coupling generative process for existing deep topic models, which incorporates skip connections into the generation of documents, enforcing strong links between the document and its multi-layer latent representations. After that, utilizing data augmentation techniques, we reformulate the deep-coupling generative process as a Markov decision process and develop a corresponding Policy Gradient (PG) based training algorithm, which can further alleviate the information reduction at higher layers. Extensive experiments demonstrate that our developed methods can effectively alleviate "*posterior collapse*" in deep topic models, contributing to providing higher-quality latent document representations.

1 Introduction

Topic modeling has become a successful technique for text analysis and been widely applied to various problems in machine learning (ML) [1, 2] and natural language processing (NLP) [3, 4] over the past two decades. Representing documents as bag-of-words (BoW) vectors, vanilla probabilistic topic models (PTMs), with latent Dirichlet allocation (LDA) [5] being the best known representative, typically formulate each document as a mixture over latent topics, where each topic is characterized by a distribution over the terms of the vocabulary and describes an interpretable semantic concept. Although being widely used, the modeling capability of these shallow topic models is still restricted by their single-layer structure, and has difficulty in exploring hierarchical thematic structures. To this end, a series of deep topic models [6, 7, 8] have been developed to extract multi-layer document representations from a text corpus, providing a more intuitive way for users to understand text data.

Recently, benefiting from the development of deep neural networks (DNNs), there has been an emerging research interest to develop neural topic models (NTMs) to boost the performance, efficiency, and usability of topic modeling with DNNs. Specifically, following the framework of variational autoencoder (VAE) [9], most NTMs [10, 11, 12] construct a variational inference network (encoder) to project each document into its stochastic latent representation, and then reconstruct the corresponding BoW observation with a stochastic/deterministic decoder. By modeling the inference/generative process with DNNs, these NTMs are more flexible and scalable than traditional Bayesian PTMs, contributing to performing large-scale downstream tasks, especially in NLP tasks [13, 14].

"*Posterior collapse*" has been widely criticized in the field of generative model [15, 16], and the occurrence of this phenomenon will cause the approximated posterior $q_\phi(\boldsymbol{z}|\boldsymbol{x})$ collapses it non-information prior distribution $p_\theta(\boldsymbol{z})$, leading their KL divergence to be colse to zero [9, 15, 16, 17].

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

For deep topic models, despite achieving attractive performance, existing PTMs or NTMs still suffer from different degrees of "*posterior collapse*", which causes their exhibiting similar or meaningless patterns at higher layers [18, 19, 20]. Although there have been several deep NTMs [19, 20] trying to alleviate this issue by constructing more flexible inference networks, the collapse phenomenon in deep NTMs may not be solved in essence, because the true posterior provided by the generative model and the objective function for optimization remain almost unchanged [17].

To extract higher-quality hierarchical latent document representations, in this paper, we develop a deep-coupling generative process equipped with a Policy Gradients (PG) based training algorithm for existing deep topic models. The main contributions of this work are as follows:

- We develop a deep-coupling generative process for deep topic models, which incorporates skip connections into the generation of documents to alleviate "*posterior collapse*".
- We take a specific NTM as an example to explain how to construct a deep topic model with the deep coupling generation process, and develop a deep-coupling hierarchical Embedding Topic Model (dc-ETM), which can be extended to other deep topic models.
- Utilizing the property of sequence-like generation process, we design a PG-based training algorithm for dc-ETM, which can further alleviate the information reduction at higher layers.
- Compared to existing deep topic models, extensive experimental results show that dc-ETMs can lead to less "*posterior collapse*" and provide higher-quality latent representations.

2 Related Work

Probabilistic Topic Model: Deep PTMs [6, 7, 8, 21, 22] are developed to infer multi-layer document representations, whose adjacent layers are connected with specific factorization. For instance, gamma belief network (GBN) [7] is constructed via factorizing the shape parameters of the gamma distributed latent representations; DPFA [6] extends PFA [23] into a multi-layer version but is restricted to model binary topic usage patterns; DirBN [8] is developed via factorizing the Dirichlet distributed topic matrix. Although providing readily interpretable multi-layer latent document representations, the representation capability of these deep PTMs is limited by adopting CRT distribution to upward propagate data information to higher layers with their backbones [18].

Neural Topic Model: Most existing NTMs [11, 19, 24, 20] can be viewed as extensions of PTMs under the VAE framework and focus on modeling the generative/inference process with DNNs. For instance, one popular research direction of NTMs is to develop more flexible inference network with reparametrization tricks [11, 19] and the other could be incorporating word embeddings into the generative model [24, 20]. However, as far as we know, few efforts have been made to alleviate the phenomenon of "*posterior collapse*" in NTMs by modifying its generative process, which is a great challenge under the framework of topic modeling and also the main contribution of this work.

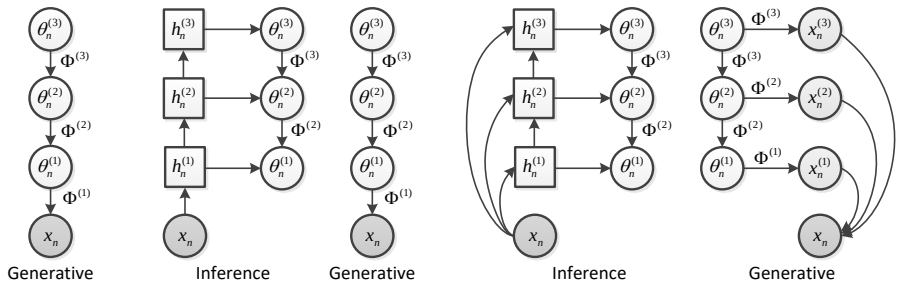
Besides, distinct from the way of combining reinforcement learning (RL) with topic models in previous works [25, 26, 27], our work is the first to formulate the topic modeling generative process as a sequential decision making one to incorporate RL-based training algorithms, which focuses on providing higher-quality latent document representations by alleviating "posterior collapse".

3 Deep-Coupling Generative Process for Deep Topic Models

To give an intuitive insight on "*posterior collapse*" in deep topic models, we provide a detailed introduction for "*posterior collapse*" in Appendix and then visualize the higher-level topics learned by a recent popular NTM named SawETM [20] in Fig. 3, which exhibit similar semantic patterns and limit its representation capability. Then, we take SawETM as an example, but not limited to this, to illustrate how to construct a deep topic model with the deep coupling generation process, leading to a novel dc-ETM in Fig. 1(c). Compared to the usual structures of deep PTMs and NTMs shown in Fig. 1(a) and 1(b), besides the design of inference network, the main difference of dc-ETM is incorporating skip connections into the generation of documents, enforcing strong links between the document and its multi-layer latent representations to alleviate "*posterior collapse*".

3.1 Deep-Coupling Hierarchical Embedding Topic Model

As a usual VAE-like model, the developed dc-ETM consists of a generative model (decoder) and an inference network (encoder). Below, we focus on presenting the generative model of dc-ETM, which



(a) Deep PTM (b) Deep NTM (c) dc-ETM

Figure 1: The overview of the network structure of (a) deep PTM, (b) deep NTM, and (c) dc-ETM developed in this paper, where the symbol definitions are consistent with those in Sec. 3.1.

can be flexibly applied for other deep topic models to alleviate "*posterior collapse*", and leave the details of the inference network to Appendix A.

Generative Model: Given a text corpus consisting of N documents $\boldsymbol{X}=\{\boldsymbol{x}_n\}_{n=1}^N$, each document can be represented as a high-dimensional sparse BoW vector $\boldsymbol{x}_n \in \mathbb{Z}^{K^{(0)}}$, where $\mathbb{Z}=\{0,1,...\}$ and $K^{(0)}$ denotes the vocabulary size. Then, from top to bottom, the generative model of the dc-ETM with L hidden layers can be formulated as

$$\boldsymbol{\theta}_n^{(l)} \sim \text{Gam}(\boldsymbol{\Phi}^{(l+1)}\boldsymbol{\theta}_n^{(l+1)}, 1/c_n^{(l+1)}), l=1,...,L-1, \cdots, \boldsymbol{\theta}_n^{(L)} \sim \text{Gam}(\boldsymbol{r}, 1/c_n^{(L+1)}), \tag{1}$$

$$\boldsymbol{x}_n \sim \text{Pois}(\sum_{l=1}^{L} \alpha^{(l)}\hat{\boldsymbol{\Phi}}^{(l)}\boldsymbol{\theta}_n^{(l)}),\ \boldsymbol{\alpha}=\text{Softmax}(\boldsymbol{\xi}),\ \boldsymbol{\phi}_k^{(l)}=\text{Softmax}(\boldsymbol{\beta}^{(l-1)^T}\boldsymbol{\beta}_k^{(l)}), l=1,...,L-1,$$

where, $\boldsymbol{\Phi}^{(l)} \in \mathbb{R}_+^{K^{(l-1)}\times K^{(l)}}$ denotes the topic matrix (factor loading) and each column $\boldsymbol{\phi}_k^{(l)} \in \mathbb{R}_+^{K^{(l-1)}}$ indicates a specific topic (factor) at layer l; $\boldsymbol{\theta}_n^{(l)} \in \mathbb{R}_+^{K^{(l)}}$ denotes the gamma distributed latent representation (topic proportions) at layer l, $K^{(l)}$ denotes the number of hidden units (topics) at layer l. Under the Poisson likelihood, the observed multivariate count vector $\boldsymbol{x}_n$ is first factorized into L equal-size latent matrix $\{\alpha^{(l)}\hat{\boldsymbol{\Phi}}^{(l)}\boldsymbol{\theta}_n^{(l)}\}_{l=1}^L$, where, $\hat{\boldsymbol{\Phi}}^{(l)} \in \mathbb{R}_+^{K^{(0)}\times K^{(l)}}$ can be regarded as the projection of topic matrix $\boldsymbol{\Phi}^{(l)}$ to the observation space and the detailed definition will be discussed in the next paragraph; $\alpha^{(l)}$ denotes the importance weight of $\hat{\boldsymbol{\Phi}}^{(l)}\boldsymbol{\theta}_n^{(l)}$ for generating the observation $\boldsymbol{x}_n$, and the summation of the whole weight vector $\boldsymbol{\alpha} \in \mathbb{R}_+^L$ is constrained to be equal to one with a Softmax normalization. Then, the latent representation $\boldsymbol{\theta}_n^{(l)}$ at layer l is further factorized into the product of the topic matrix $\boldsymbol{\Phi}^{(l+1)} \in \mathbb{R}_+^{K^{(l)}\times K^{(l+1)}}$ and topic proportions $\boldsymbol{\theta}_n^{(l+1)} \in \mathbb{R}_+^{K^{(l+1)}}$ at the next layer under the shape of gamma distribution. The top layer's latent representation $\boldsymbol{\theta}_n^{(L)}$ shares the same gamma shape parameters $\boldsymbol{r} \in \mathbb{R}_+^{K^{(L)}}$ and we apply a gamma distributed prior on the scale parameters $c_n^{(l)}$ for $l \in \{2,...,L+1\}$. With the recent popular distributed topic representation in NTMs [24, 28], each topic $\boldsymbol{\phi}_k^{(l)}$ is treated as the result of applying a Softmax normalization on the inner product of its distributed representation $\boldsymbol{\beta}_k^{(l)} \in \mathbb{R}^D$ and topic embedding matrix $\boldsymbol{\beta}^{(l-1)} \in \mathbb{R}^{D\times K^{(l-1)}}$ at the previous layer, where D denotes the dimension of the embedding space.

The projections of topic matrices to the observation space, denoted as $\{\hat{\boldsymbol{\Phi}}^{(l)}\}_{l=1}^L$, build the straightforward connections between the document $\boldsymbol{x}_n$ and its multi-layer latent representations $\{\boldsymbol{\theta}_n^{(l)}\}_{l=1}^L$, which alleviates the information reduction at higher layers by sharing the pressure of document modeling with all hidden layers. To reduce the computation and storage cost of the developed dc-ETM, we develop two variants for $\hat{\boldsymbol{\phi}}_k^{(l)} \in \mathbb{R}_+^{K^{(0)}}$ without introducing any extra parameter. The one variant is adopting the property of topic hierarchy elaborated in Sec. 3.2 to obtain each $\hat{\boldsymbol{\phi}}_k^{(l)}$ by successively multiplying topic matrices at lower layers as

$$\hat{\boldsymbol{\phi}}_k^{(l)} = \prod_{t=1}^{l-1} \boldsymbol{\Phi}^{(t)}\boldsymbol{\phi}_k^{(l)}, \tag{2}$$

and the other variant is treating the projection $\hat{\boldsymbol{\phi}}_k^{(l)}$ as the result of the inner product of its distributed representation $\boldsymbol{\beta}_k^{(l)}$ and the word embedding matrix $\boldsymbol{\beta}^{(0)}$ at the observed space, as follows

$$\hat{\boldsymbol{\phi}}_k^{(l)} = \text{Softmax}({\boldsymbol{\beta}^{(0)}}^T \boldsymbol{\beta}_k^{(l)}). \tag{3}$$

We emphasize that the first variant can be used to extend most existing deep topic models, while the latter is limited to NTMs equipped with topic embedding techniques. We use the suffix $-\alpha$ and $-\beta$ to distinguish the variants defined in Eq. (2) and Eq. (3), and their detailed implementations can be found in Appendix I.

Generally speaking, the deep-coupling generative process in dc-ETM not only preserves the hierarchy of traditional deep topic models, leading to multi-layer document representations to enhance the modeling capability and interpretability, but also alleviates the issue that the amount of information will decrease rapidly with the network going deeper, benefiting from building the straightforward connections between observation $\boldsymbol{x}_n$ and its higher-level latent representations $\{\boldsymbol{\theta}_n^{(l)}\}_{l>1}$. Besides alleviating "*posterior collapse*", the characteristics of deep-coupling network structure of dc-ETM also brings us a new view to design the corresponding inference network and training algorithm.

Inference Network: The details of the inference network of dc-ETM can be found in Appendix A.

3.2 Model Property

Sequence-like Generative Process: Taking advantages of the properties of the Poisson distribution, the original generative process of the observed data $\boldsymbol{x}_n$ defined in Eq. (1) can be rewritten as:

$$\boldsymbol{x}_n = \sum_{l=1}^{L} \boldsymbol{x}_n^{(l)},\ \boldsymbol{x}_n^{(l)} \sim \text{Pois}(\alpha^{(l)} \hat{\boldsymbol{\Phi}}^{(l)} \boldsymbol{\theta}_n^{(l)}), \tag{4}$$

where $\boldsymbol{x}_n^{(l)}$ denotes the augmented observation at layer l, and is generated from the Poisson distribution with a rate of $\alpha^{(l)} \hat{\boldsymbol{\Phi}}^{(l)} \boldsymbol{\theta}_n^{(l)}$. Then, the observed data $\boldsymbol{x}_n$ can be regarded as not only the summation over these augmented vectors $\{\boldsymbol{x}_n^{(l)}\}_{l=1}^{L}$, but also equal to the weighted summation over the latent vectors $\{\hat{\boldsymbol{\Phi}}^{(l)} \boldsymbol{\theta}_n^{(l)}\}_{l=1}^{L}$ on the mean, where the weight vector $\boldsymbol{\alpha}$ satisfies the constraint $\sum_{l=1}^{L} \alpha^{(l)} = 1$. Rethinking the generative process of the developed dc-ETM reformulated in Eq. (4), the set of augmented observation vectors $\{\boldsymbol{x}_n^{(l)}\}_{l=1}^{L}$ can naturally form an observation sequence $[\boldsymbol{x}_n^{(L)}, ..., \boldsymbol{x}_n^{(1)}]$ by sorting these vectors according to their dependencies in the generative process (from deep to shallow). For each hidden layer (time step) l, the generative process will first incorporate the prior information passing from deeper hidden layers $\{\boldsymbol{\theta}_n^{(t)}\}_{t>l}$, and then generate the latent representation $\boldsymbol{\theta}_n^{(l)}$ at the current layer (time step), which not only is supposed to generate the current observation vector $\boldsymbol{x}_n^{(l)}$ under the Poisson likelihood, but also introduces the information into the shape parameter of the following gamma distributed latent representation $\boldsymbol{\theta}_n^{(l-1)}$ at the next layer (time step).

Thus, the deep-coupling generative process of dc-ETM originally defined in Eq. (1) can be naturally reinterpreted from the perspective of sequence generation, and its reformulation defined in Eq. (4) can be also equivalently reformulated as:

$$\boldsymbol{x}_n \sim \sum_{l=1}^{L} \text{Pois}(\alpha^{(l)} \hat{\boldsymbol{\Phi}}^{(l)} \boldsymbol{\theta}_n^{(l)}), \tag{5}$$

providing an intuitive insight for the decomposition of the likelihood function in Sec. 4.1.

Hierarchical Semantic Topics: The developed dc-ETM can naturally interpret each semantic topic $\boldsymbol{\phi}_k^{(l)}$ at layer l by visualizing its projection to the vocabulary space calculated as $\{[\prod_{t=1}^{l-1} \boldsymbol{\Phi}^{(t)}] \boldsymbol{\phi}_k^{(l)}\}_{k=1}^{K^{(l)}}$, and each document can also be roughly seen as a random mixture over $K^{(l)}$ topics with $\boldsymbol{\theta}_n^{(l)}$ being the corresponding topic proportions at layer l as

$$\mathbb{E}\left[\boldsymbol{x}_n | \boldsymbol{\theta}_n^{(l)}, \left\{\boldsymbol{\Phi}^{(t)}, c_n^{(t)}\right\}_{t=1}^{l}\right] = \left[\prod_{t=1}^{l} \boldsymbol{\Phi}^{(t)}\right] \frac{\boldsymbol{\theta}_n^{(l)}}{\prod_{t=2}^{l} c_n^{(t)}}, \tag{6}$$

which can be obtained with the law of total expectation. Moreover, similar to the underlying idea of the deep learning, the topics learned by dc-ETM tend to be more specific at lower (bottom) layers and those at higher (top) layers are more general, as shown in Fig. 7.

Secondly, in dc-ETM, both words $\boldsymbol{\beta}^{(0)} \in \mathbb{R}^{D\times K^{(0)}}$ and hierarchical topics $\{\boldsymbol{\beta}^{(l)} \in \mathbb{R}^{D\times K^{(l)}}\}_{l=1}^{L}$ are represented as embedding vectors under the same semantic space, contributing to intuitively measuring and visualizing the distance between different topics (words), which has been proven to be effective in capturing the underlying semantic structure as shown in Fig. 5(a).

4 Policy Gradient based Training Algorithm

4.1 ELBO of dc-ETM

As a VAE-like NTM, the developed dc-ETM can be trained like usual VAEs by directly maximizing the evidence lower bound (ELBO), specifically as

$$L(\boldsymbol{x}_n) = \mathbb{E}_{q(\boldsymbol{\theta}_n|\boldsymbol{x}_n)}[\ln p(\boldsymbol{x}_n|\boldsymbol{\theta}_n)] - \mathrm{KL}(q(\boldsymbol{\theta}_n|\boldsymbol{x}_n)||p(\boldsymbol{\theta}_n)), \tag{7}$$

where the first term is the expected log-likelihood and the other term is the Kullback–Leibler (KL) divergence from the prior $p(\boldsymbol{\theta}_n)$ to the variational posterior $q(\boldsymbol{\theta}_n|\boldsymbol{x}_n)$.

Through introducing the augmented vectors $\{\boldsymbol{x}_n^{(l)}\}_{l=1}^{L}$, the log-likelihood of $\boldsymbol{x}_n$ in dc-ETM can be equivalently reformulated as

$$\begin{aligned}\ln p(\boldsymbol{x}_n|\boldsymbol{\theta}_n) &= \mathbb{E}_{q(\{\boldsymbol{x}_n^{(l)}\}_{l=1}^{L}|-)}\left[\ln p(\boldsymbol{x}_n|\{\boldsymbol{x}_n^{(l)}\}_{l=1}^{L})\prod_{l=1}^{L} p(\boldsymbol{x}_n^{(l)}|\boldsymbol{\theta}_n^{(l)})\right] \\ &= \mathbb{E}_{q(\{\boldsymbol{x}_n^{(l)}\}_{l=1}^{L}|-)}\left[\ln p(\boldsymbol{x}_n|\{\boldsymbol{x}_n^{(l)}\}_{l=1}^{L})\right] + \mathbb{E}_{q(\{\boldsymbol{x}_n^{(l)}\}_{l=1}^{L}|-)}\left[\sum_{l=1}^{L}\ln p(\boldsymbol{x}_n^{(l)}|\boldsymbol{\theta}_n^{(l)})\right],\end{aligned} \tag{8}$$

where the function in the second expectation term can be treated as the summation of the set of log-likelihood of $\{\boldsymbol{x}_n^{(l)}\}_{l=1}^{L}$. Due to the hierarchical network structure, the KL divergence term can be factorized as

$$\mathrm{KL}(q(\boldsymbol{\theta}_n|\boldsymbol{x}_n)||p(\boldsymbol{\theta}_n)) = \sum_{l=1}^{L}\mathbb{E}_{q(\boldsymbol{\theta}_n^{(l)}|-)}\left[\ln\frac{q(\boldsymbol{\theta}_n^{(l)}|-)}{p(\boldsymbol{\theta}_n^{(l)}|\boldsymbol{\Phi}^{(l+1)},\boldsymbol{\theta}_n^{(l+1)})}\right], \tag{9}$$

where $q(\boldsymbol{\theta}_n^{(l)}|-)$ is constructed by a Weibull-based inference network described in Appendix. A and $p(\boldsymbol{\theta}_n^{(l)}|\boldsymbol{\Phi}^{(l+1)},\boldsymbol{\theta}_n{}^{(l+1)})$ satisfies a gamma prior in Eq. (1), and their KL divergence has an analytic expression, benefiting from adopting the Weibull reparameterization technique [19].

Combining the aforementioned derivations, the ELBO of dc-ETM can be equivalently rewritten as

$$\begin{aligned}L(\boldsymbol{x}_n) =& \mathbb{E}_{q(\{\boldsymbol{x}_n^{(l)}\}_{l=1}^{L}|-)}\left[\ln p(\boldsymbol{x}_n|\{\boldsymbol{x}_n^{(l)}\}_{l=1}^{L})\right] + \mathbb{E}_{q(\{\boldsymbol{x}_n^{(l)},\boldsymbol{\theta}_n^{(l)}\}_{l=1}^{L}|-)}\left[\sum_{l=1}^{L}\ln p(\boldsymbol{x}_n^{(l)}|\boldsymbol{\theta}_n^{(l)})\right] \\ &- \sum_{l=1}^{L}\mathbb{E}_{q(\boldsymbol{\theta}_n^{(l)}|-)}\left[\ln\frac{q(\boldsymbol{\theta}_n^{(l)}|-)}{p(\boldsymbol{\theta}_n^{(l)}|\boldsymbol{\Phi}^{(l+1)},\boldsymbol{\theta}_n^{(l+1)})}\right],\end{aligned} \tag{10}$$

which can be directly optimized with gradient-based methods to update both the encoder parameters $\boldsymbol{\Omega}$ and decoder parameters $\boldsymbol{\Psi}$ in dc-ETM. We emphasize that, after deriving the augmented vectors $\{\boldsymbol{x}_n^{(l)}\}_{l=1}^{L}$ from $\boldsymbol{x}_n$ via data augmentation technique [7], the first expectation term in $L(\boldsymbol{x}_n)$ will be a constant and the ELBO can be directly optimized by maximizing the following loss function

$$\begin{aligned}\hat{L}(\{\boldsymbol{x}_n^{(l)}\}_{l=1}^{L}) &= \sum_{l=1}^{L}\mathbb{E}_{q(\boldsymbol{\theta}_n^{(l)}|-)}\left[\ln p(\boldsymbol{x}_n^{(l)}|\boldsymbol{\theta}_n^{(l)})\right] - \sum_{l=1}^{L}\mathbb{E}_{q(\boldsymbol{\theta}_n^{(l)}|-)}\left[\ln\frac{q(\boldsymbol{\theta}_n^{(l)}|-)}{p(\boldsymbol{\theta}_n^{(l)}|\boldsymbol{\Phi}^{(l+1)},\boldsymbol{\theta}_n^{(l+1)})}\right], \\ &= \sum_{l=1}^{L}\mathbb{E}_{q(\boldsymbol{\theta}_n^{(l)}|-)}\left[\ln\frac{p(\boldsymbol{x}_n^{(l)}|\boldsymbol{\theta}_n^{(l)})p(\boldsymbol{\theta}_n^{(l)}|\boldsymbol{\Phi}^{(l+1)},\boldsymbol{\theta}_n^{(l+1)})}{q(\boldsymbol{\theta}_n^{(l)}|-)}\right], \\ &= \sum_{l=1}^{L}\hat{L}^{(l)}(\boldsymbol{x}_n^{(l)};\alpha^{(l)},\hat{\boldsymbol{\Phi}}^{(l)},\boldsymbol{\theta}_n^{(l)},\{\boldsymbol{\Phi}^{(t)},\boldsymbol{\theta}_n^{(t)}\}_{t>l})\end{aligned} \tag{11}$$

which can be roughly treated as the ELBO of a sequence $[\boldsymbol{x}_n^{(L)}, ..., \boldsymbol{x}_n^{(1)}]$ generated from a sequence of latent representations $[\boldsymbol{\theta}_n^{(L)}, ..., \boldsymbol{\theta}_n^{(1)}]$ [29, 30], and naturally meets the sequence-like generative process of dc-ETM as discussed in Sec. 3.2.

4.2 Optimization with Policy Gradient

Similar to RNN-based model, after augmenting $\{\boldsymbol{x}_n^{(l)}\}_{l=1}^{L}$ from $\boldsymbol{x}_n$, the loss function of dc-ETM defined in Eq. (11) is equal to the summation of L sub-loss functions, where each sub-loss function $\hat{L}^{(l)}(\boldsymbol{x}_n^{(l)})$ can be equivalently regarded as a separate loss of a subsequence generation model that is only a part of the whole sequential generative model and expected to output $\boldsymbol{x}_n^{(l)}$ at the final time step l. Inspired by the great success achieved by RL methods [31, 32, 33, 34] in learning a stable long sequence (Markov decision process) with high quality, we consider the sequence-like generation procedure of a L-layer dc-ETM as a Markov decision process with L time steps, and develop a novel training mechanism based on Policy Gradient [33] for dc-ETM, which injects the future rewards obtained from generating the suffix subsequence into each current sub-loss function $\hat{L}^{(l)}(\boldsymbol{x}_n^{(l)})$.

Specifically, we treat the whole dc-ETM as a stochastic policy network $\pi(a_n^{(l)}|s_n^{(l)})$ expected to generate a fixed-length action sequence $[a_n^{(L)}, ..., a_n^{(1)}]$ from the observation $\boldsymbol{x}_n$, defining the state $s_n^{(l)}$ as $\{\boldsymbol{x}_n, \{\boldsymbol{\Phi}^{(t)}, \boldsymbol{\theta}_n^{(t)}\}_{t>l}\}$ and the action $a_n^{(l)}$ as $\alpha^{(l)}\hat{\boldsymbol{\Phi}}^{(l)}\boldsymbol{\theta}_n^{(l)}$. For each time step l, given the current state $s_n^{(l)}$, the policy network $\pi(a_n^{(l)}|s_n^{(l)})$ will first sample $\boldsymbol{\theta}_n^{(l)}$ from the inference network via

$$\boldsymbol{\theta}_n^{(l)} \sim q(\boldsymbol{\theta}_n^{(l)}|\boldsymbol{x}_n, \{\boldsymbol{\Phi}^{(t)}, \boldsymbol{\theta}_n^{(t)}\}_{t>l}), \tag{12}$$

and further obtain the corresponding action as

$$a_n^{(l)} = \alpha^{(l)}\hat{\boldsymbol{\Phi}}^{(l)}\boldsymbol{\theta}_n^{(l)}, \tag{13}$$

which can be regarded as directly drawing from $\pi(a_n^{(l)}|s_n^{(l)})$. The state transition is deterministic after an action has been chosen, indicating that the next state $s_n^{(l-1)} = \{\boldsymbol{x}_n, \{\boldsymbol{\Phi}^{(t)}, \boldsymbol{\theta}_n^{(t)}\}_{t>l-1}\}$ if the current state $s_n^{(l)} = \{\boldsymbol{x}_n, \{\boldsymbol{\Phi}^{(t)}, \boldsymbol{\theta}_n^{(t)}\}_{t>l}\}$ and the action $a_n^{(l)} = \alpha^{(l)}\hat{\boldsymbol{\Phi}}^{(l)}\boldsymbol{\theta}_n^{(l)}$.

Then we take the separate loss $\hat{L}^{(l)}(\boldsymbol{x}_n^{(l)})$ defined in Eq. (11) as the immediate reward at the the time step l, formulated as

$$r(s_n^{(l)}, a_n^{(l)}) = \mathbb{E}_{q(\boldsymbol{\theta}_n^{(l)}|-)}\left[\ln p(\boldsymbol{x}_n^{(l)}|\alpha^{(l)}, \hat{\boldsymbol{\Phi}}^{(l)}, \boldsymbol{\theta}_n^{(l)})\right] - \mathbb{E}_{q(\boldsymbol{\theta}_n^{(l)}|-)}\left[\ln \frac{q(\boldsymbol{\theta}_n^{(l)}|-)}{p(\boldsymbol{\theta}_n^{(l)}|\boldsymbol{\Phi}^{(l+1)}, \boldsymbol{\theta}_n^{(l+1)})}\right], \tag{14}$$

and the action-value function can be formulated as

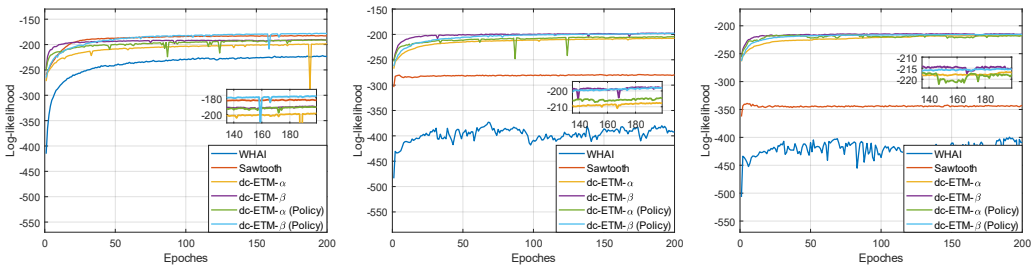
$$Q^{\pi}(s_n^{(l)}, a_n^{(l)}) = r(s_n^{(l)}, a_n^{(l)}) + \mathbb{E}_{\pi}\left[\sum_{i=1}^{l-1}\gamma^i r(s_n^{(l-i)}, a_n^{(l-i)})\right], \tag{15}$$

which indicates the expected accumulative reward starting from state $s_n^{(l)}$, taking action $a_n^{(l)}$, and then generating the suffix subsequence $[a_n^{(l-1)}, ..., a_n^{(1)}]$ with the policy network $\pi(a_n^{(l)}|s_n^{(l)})$ and the discount factor $0 < \gamma \leq 1$.

Following [31], the objective function of training dc-ETM with policy gradient can be estimated (on one episode) as

$$J(\boldsymbol{x}_n; \boldsymbol{\Omega}, \boldsymbol{\Psi}) \simeq \sum\nolimits_{l=1}^{L}\int_{a_n^{(l)}} \pi(a_n^{(l)}|s_n^{(l)})Q^{\pi}(s_n^{(l)}, a_n^{(l)}) = \sum\nolimits_{l=1}^{L}\mathbb{E}_{\pi(a_n^{(l)}|s_n^{(l)})}\left[Q^{\pi}(s_n^{(l)}, a_n^{(l)})\right], \tag{16}$$

where $\boldsymbol{\Omega}$ and $\boldsymbol{\Psi}$ indicate the encoder and decoder parameters in dc-ETM respectively. We note that the expectation $\mathbb{E}[\cdot]$ can be approximated by sampling methods based on the Weibull reparameterization, and the objective function can be directly optimized by advanced gradient descent algorithms, like Adam [35] and RMSprop [36]. We provide the details of PG-based training algorithm in Appendix C.



(a) $\ln p(\boldsymbol{x}_n|\hat{\boldsymbol{\Phi}}^{(1)}\boldsymbol{\theta}_n^{(1)})$ (b) $\ln p(\boldsymbol{x}_n|\hat{\boldsymbol{\Phi}}^{(3)}\boldsymbol{\theta}_n^{(3)})$ (c) $\ln p(\boldsymbol{x}_n|\hat{\boldsymbol{\Phi}}^{(5)}\boldsymbol{\theta}_n^{(5)})$

Figure 2: Point log-likelihood $\ln p(\boldsymbol{x}_n|\hat{\boldsymbol{\Phi}}^{(t)}\boldsymbol{\theta}_n^{(t)})$ of different deep topic models on 20News dataset as a function of iterative epoches, where $\hat{\boldsymbol{\Phi}}^{(t)}\boldsymbol{\theta}_n^{(t)}$ can be treated as the projection of $\boldsymbol{\theta}_n^{(t)}$ from the latent space to the observation space, as discussed in Sec. 3.2.

5 Experiments

To evaluate the effectiveness of the developed dc-ETM and the corresponding policy gradients (PG) based training algorithm, we make extensive experiments on both quantitative and qualitative aspects. Considering there are two dc-ETM variants as described in 3.1, we use the suffix $-\alpha$ and $-\beta$ to distinguish the variants defined in Eq. (2) and Eq. (3) respectively, and highlight whether the dc-ETM is trained with the PG based algorithm in Sec. 4.2.

5.1 Datasets and Baselines

Datasets: Four widely used document benchmarks, specifically **R8** [37], 20Newsgroups (**20News**) [38], Reuters Corpus Volume I (**RCV1**) [39] and World Wide Web Knowledge Base (**WebKB**) [40] are included in the following experiments. We summarize the statistics of benchmarks in Appendix D and follow the procedure in [20] to preprocess these documents to obtain their BoW representations.

Baselines: We compare the developed dc-ETMs with a series of topic models, which can be roughly divided into two categories: 1) shallow topic models such as LDA [5], AVITM [11] and ETM [24], where LDA is a PTM and the others are NTMs; 2) deep topic models including PGBN [7], WHAI [19] and SawETM [20], where PGBN is a deep PTM and the others are deep NTMs. We emphasize that WHAI and SawETM are the most relevant strong baselines for comparison, both of which provide hierarchical Weibull-based latent document representations, and SawETM has achieved state-of-the-art performance on unsupervised document modeling and clustering tasks.

Experimental Settings: To make a fair comparison, we set the same network structure for all deep topic models as $[256, 128, 64, 32, 16]$ from shallow to deep. For PTMs, we use the default hyperparameter settings in their published papers and accelerate the Gibbs sampling with GPU. For NTMs, we set the size of their hidden layers as 256, the embedding size as 100 for them incorporating word embeddings, like ETM, SawETM and dc-ETMs, and the mini-batch size as 200. For optimization, we adopt the same Adam optimizer [41] with a learning rate of 1e-2. All experiments are performed with an Nvidia RTX 3090 GPU and implemented with PyTorch [42].

5.2 Quantitative Comparisons

To investigate whether the proposed skip-connection structures in both dc-ETM variants can alleviate "*posterior collapse*", especially at higher layers, we compare them with other popular deep topic models in the first part. Then, to investigate whether the mitigation of "*posterior collapse*" can improve the quality of latent document representations at higher layers, we conduct more quantitative comparisons in the rest parts. We report the error bars in Appendix E.

Document Modeling: In Fig. 2, for each deep topic model, we plot the curve of point log-likelihood $\ln p(\boldsymbol{x}_n|-)$ as a function of iterative epochs conditioned on the t-th-layer reconstruction $\hat{\boldsymbol{\Phi}}^{(t)}\boldsymbol{\theta}_n^{(t)}$, which can be used to measure the degree of "*posterior collapse*" by the relevance between the data sample $\boldsymbol{x}_n$ and its latent representation $\boldsymbol{\theta}_n^{(t)}$. From the results, we can see that although SawETM and WHAI achieve a comparable performance with dc-ETMs on the first hidden layer in Fig. 2(a), their reconstruction quality decreases dramatically with the network going deeper in Fig. 2(b) and 2(c), potentially reflecting that little data information can be propagated to higher layers of these

traditional deep topic models. Benefiting from introducing skip connections into the generative process, dc-ETMs can significantly alleviate "*posterior collapse*" at higher hidden layers.

Perplexity & Topic Diversity: To make a more comprehensive quantitative comparison, we use the average of heldout-word perplexities (the lower is the better) and topic diversities (the higher is the better) across all hidden layers to measure the document modeling performance and topic quality of these deep topic models with $\{\boldsymbol{\theta}_n^{(t)}\}_{t=1}^T$ and $\{\boldsymbol{\Phi}_n^{(t)}\}_{t=1}^T$, respectively. The experimental settings are consistent with those in [20], and the experimental results have been exhibited in Table 3. Benefiting from hierarchical network structures, the modeling capability of deep topic models generally outperform those shallow ones. Thanks to enhancing the connections between the observation and multiple hidden layers with the deep-coupling generative process, the developed dc-ETMs achieve lower perplexity scores and provide higher-quality topics than traditional topic models. Then, the PG-based training algorithm brings further performance improvement to our dc-ETMs.

Table 1: Comparisons of the average of perplexities and topic diversities across all hidden layers on various benchmarks.

Model	Perplexity			Topic Diversity		
	R8	20News	RCV1	R8	20News	RCV1
LDA [5]	996	1091	1242	0.288	0.356	0.423
AVITM [11]	561	1030	1121	0.330	0.408	0.483
ETM [24]	985	989	1480	0.352	0.410	0.524
PGBN [7]	657	743	1086	0.221	0.186	0.355
WHAI [19]	773	870	1192	0.183	0.158	0.294
SawETM [20]	530	732	920	0.207	0.175	0.331
dc-ETM-α	521	730	912	0.212	0.281	0.435
dc-ETM-β	427	710	873	0.346	0.429	0.566
dc-ETM-α (Policy)	463	707	896	0.279	0.385	0.519
dc-ETM-β (Policy)	**420**	**647**	**841**	**0.379**	**0.456**	**0.584**

Document Clustering: To evaluate the quality of the extracted latent document representations on downstream tasks, we consider document clustering, where we use the topic models after training to extract the latent representations of the testing documents and then use k-means to predict the clustering labels. Using the Purity and Normalized Mutual Information (NMI) as metrics (the higher the better), the results shown in Table 4 demonstrate that concatenating hierarchical latent document representations extracted by traditional deep topic models cannot improve and even hurt the clustering performance, potentially indicating that the latent representations at higher layers are meaningless. However, distinct from traditional deep topic models, the concatenation operation on the latent representations of dc-ETMs can significantly improve the performance, which can be attributed to enforcing strong links between the multi-layer representations and the observation with the skip connections in the generation.

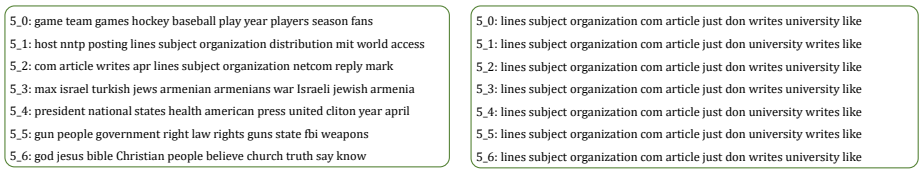
Table 2: Document clustering comparison on the 1st hidden layer or the concatenation of all hidden layers of different topic models.

Model	Layer	WebKB		20News		R8	
		Purity	NMI	Purity	NMI	Purity	NMI
LDA	1	53.40	11.23	41.79	45.15	65.74	40.47
AVITM	1	54.18	17.77	42.33	46.33	70.96	41.20
ETM	1	51.43	12.52	42.61	48.40	72.20	41.28
PGBN	1	55.37	16.27	43.30	46.51	74.52	41.24
	All	53.58	15.39	41.17	44.20	72.93	31.35
WHAI	1	59.89	25.95	42.25	46.98	74.70	43.98
	All	57.46	24.49	32.00	37.51	70.80	41.25
SawETM	1	57.89	21.91	43.33	50.77	75.25	42.97
	All	51.75	20.60	38.69	39.33	75.89	39.55
dc-ETM-α	1	61.14	26.29	32.81	43.64	75.60	39.83
	All	63.18	28.35	41.83	44.52	76.31	43.73
dc-ETM-β	1	54.71	21.43	39.80	44.30	74.30	38.63
	All	67.29	33.60	45.00	46.20	76.25	45.64
dc-ETM-α (Policy)	1	49.71	14.86	37.88	43.56	71.65	32.73
	All	64.32	33.65	42.21	45.59	77.46	44.60
dc-ETM-β (Policy)	1	57.32	26.05	40.11	44.12	71.30	38.34
	All	**69.32**	**38.53**	**48.60**	**55.79**	**78.29**	**48.62**

5.3 Qualitative Analysis

As discussed in Sec. 3.2, the developed dc-ETM inherits both the characteristics of hierarchical topic structure and semantic topic embeddings. Then we compare the hierarchical topics of a 5-layer dc-ETM trained on 20News with those learned by SawETM for qualitative analysis.

Topic Visualization: With the visualization techniques [7], we exhibit the 5th-layer topics learned by dc-ETM and SawETM on 20News in Fig. 3 and Fig. 7, where each topic is interpreted by its top-10 words by sorting the word probabilities by descending order. Obviously, the topics learned by SawETM are quite similar, explaining the reason why concatenating its hierarchical latent document representations cannot improve and even hurt the performance on downstream tasks. On the contrary, the developed dc-ETM can learn meaningful and diverse topics at higher layers, indicating that more



(a) $\hat{\mathbf{\Phi}}^{(5)}$ learned by dc-ETM (b) $\hat{\mathbf{\Phi}}^{(5)}$ learned by SawETM

Figure 3: The 5th-layer topics learned by dc-ETM and SawETM with the same network structure on 20News, where each topic is interpreted by its top-10 words. More comparisons refer to Appendix F.

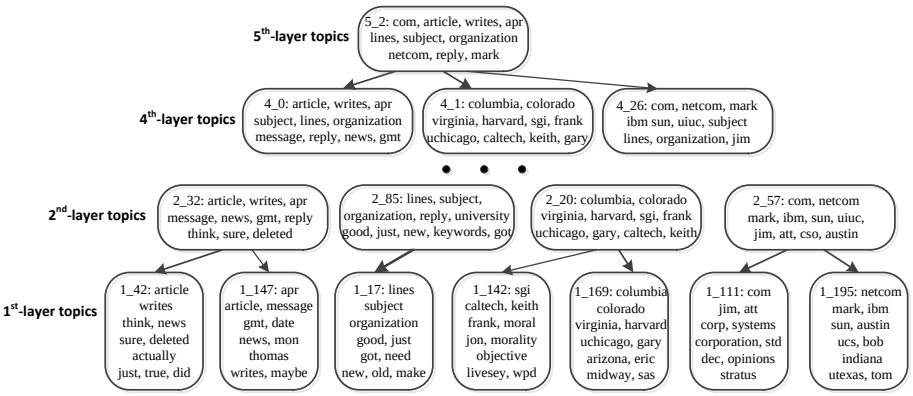
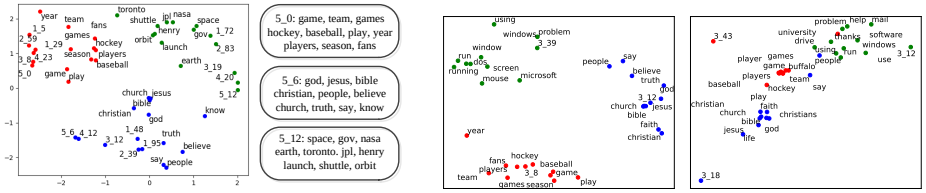


Figure 4: A hierarchical topic tree example learned by a 5-layer dc-ETM on 20News dataset.

data information is passed to higher layers to alleviate "*posterior collapse*". We also exhibit a 5-layer topic tree learned by dc-ETM in Fig. 7 to illustrate the topic hierarchy of dc-ETM in Appendix N.



(a) dc-ETM (b) dc-ETM (c) SawETM

Figure 5: t-SNE visualization of a) multiple 5-layer hierarchical topic trees learned by dc-ETM, whose leaf nodes are distinguished by different colors; b) and c) various semantic topics equipped with their own top-10 representative word embeddings learned by dc-ETM and SawETM on 20News.

Topic Embedding Visualization: After extracting hierarchical topic trees by dc-ETM, we visualize some of these trees originated from different topic nodes at layer 5 by projecting their semantic embeddings with t-SNE [43]. As shown in Fig. 5(a), we can find that the topics in the same topic tree tend to be closer than others from different trees in the semantic embedding space and similar phenomenon occurs in the words for describing the same root topic, which indicates the hierarchy learn by dc-ETM is of high quality. Note that we also visualize the topics consisting of similar top words in Fig. 5(b) and 5(c), learned by dc-ETM and SawETM respectively, which demonstrates that the developed dc-ETM can provide more meaningful and discriminative topic and word embeddings.

6 Conclusion

To provide higher-quality hierarchical latent representations for deep topic modeling, in this paper, with the deep-coupling generative process, we develop a novel dc-ETM, which is constructed by introducing skip connections into the generative process of GBN and also incorporates both topic embedding and Weibull reparameterization techniques. Utilizing the property of sequence-like generation process, we design a PG-based training algorithm for dc-ETM to further alleviate the information reduction at higher layers. We note that the main idea of designing dc-ETM equipped with the PG-based training algorithm can potentially be extended to other deep topic models.

References

[1] David M Blei and Michael I Jordan. Modeling annotated data. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 127–134, 2003.

[2] Francesca De Battisti, Alfio Ferrara, and Silvia Salini. A decade of research in statistics: A topic model approach. *Scientometrics*, 103(2):413–433, 2015.

[3] Chenghua Lin and Yulan He. Joint sentiment/topic model for sentiment analysis. In *Proceedings of the 18th ACM conference on Information and knowledge management*, pages 375–384, 2009.

[4] Jordan Boyd-Graber, David Blei, and Xiaojin Zhu. A topic model for word sense disambiguation. In *Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL)*, pages 1024–1033, 2007.

[5] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3(Jan):993–1022, 2003.

[6] Zhe Gan, Changyou Chen, Ricardo Henao, David Carlson, and Lawrence Carin. Scalable deep poisson factor analysis for topic modeling. In *International Conference on Machine Learning*, pages 1823–1832, 2015.

[7] Mingyuan Zhou, Yulai Cong, and Bo Chen. Augmentable gamma belief networks. *Journal of Machine Learning Research*, 17(1):5656–5699, 2016.

[8] He Zhao, Lan Du, Wray L. Buntine, and Mingyuan Zhou. Dirichlet belief networks for topic structure learning. In *Advances in Neural Information Processing Systems*, pages 7966–7977, 2018.

[9] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *International Conference on Learning Representations*, 2014.

[10] Yishu Miao, Lei Yu, and Phil Blunsom. Neural variational inference for text processing. In *International Conference on Machine Learning*, volume 48, pages 1727–1736, 2016.

[11] Akash Srivastava and Charles Sutton. Autoencoding variational inference for topic models. In *International Conference on Learning Representations*, 2017.

[12] Dallas Card, Chenhao Tan, and Noah A. Smith. A neural framework for generalized topic models. *CoRR*, abs/1705.09296, 2017.

[13] Hao Zhang, Bo Chen, Long Tian, Zhengjue Wang, and Mingyuan Zhou. Variational heteroencoder randomized gans for joint image-text modeling. *arXiv preprint arXiv:1905.08622*, 2019.

[14] He Zhao, Dinh Phung, Viet Huynh, Yuan Jin, Lan Du, and Wray Buntine. Topic modelling meets deep neural networks: A survey. *arXiv preprint arXiv:2103.00498*, 2021.

[15] Casper Kaae Sønderby, Tapani Raiko, Lars Maaløe, Søren Kaae Sønderby, and Ole Winther. Ladder variational autoencoders. In *Advances in Neural Information Processing Systems*, pages 3738–3746, 2016.

[16] Lars Maaløe, Marco Fraccaro, Valentin Liévin, and Ole Winther. BIVA: A very deep hierarchy of latent variables for generative modeling. In *Advances in Neural Information Processing Systems*, pages 6548–6558, 2019.

[17] Adji B Dieng, Yoon Kim, Alexander M Rush, and David M Blei. Avoiding latent variable collapse with generative skip models. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 2397–2405. PMLR, 2019.

[18] Xuhui Fan, Bin Li, Yaqiong Li, and Scott A. Sisson. Poisson-randomised dirbn: Large mutation is needed in dirichlet belief networks. In *International Conference on Machine Learning*, volume 139, pages 3068–3077, 2021.

[19] Hao Zhang, Bo Chen, Dandan Guo, and Mingyuan Zhou. WHAI: weibull hybrid autoencoding inference for deep topic modeling. In *International Conference on Learning Representations*, 2018.

[20] Zhibin Duan, Dongsheng Wang, Bo Chen, Chaojie Wang, Wenchao Chen, Yewen Li, Jie Ren, and Mingyuan Zhou. Sawtooth factorial topic embeddings guided gamma belief network. In *International Conference on Machine Learning*, volume 139, pages 2903–2913, 2021.

[21] Yulai Cong, Bo Chen, Hongwei Liu, and Mingyuan Zhou. Deep latent dirichlet allocation with topic-layer-adaptive stochastic gradient riemannian MCMC. In *International Conference on Machine Learning*, volume 70, pages 864–873, 2017.

[22] Mingyuan Zhou, Yulai Cong, and Bo Chen. The poisson gamma belief network. *Advances in Neural Information Processing Systems*, 28, 2015.

[23] Mingyuan Zhou, Lauren Hannah, David Dunson, and Lawrence Carin. Beta-negative binomial process and poisson factor analysis. In *Artificial Intelligence and Statistics*, pages 1462–1471, 2012.

[24] Adji Bousso Dieng, Francisco J. R. Ruiz, and David M. Blei. Topic modeling in embedding spaces. *Trans. Assoc. Comput. Linguistics*, 8:439–453, 2020.

[25] Lin Gui, Jia Leng, Gabriele Pergola, Yu Zhou, Ruifeng Xu, and Yulan He. Neural topic model with reinforcement learning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3478–3483, 2019.

[26] Zeinab Shahbazi and Yung-Cheol Byun. Topic modeling in short-text using non-negative matrix factorization based on deep reinforcement learning. *Journal of Intelligent & Fuzzy Systems*, 39 (1):753–770, 2020.

[27] Amit Kumar, Nazanin Esmaili, and Massimo Piccardi. A reinforced variational autoencoder topic model. In *International Conference on Neural Information Processing*, pages 360–369. Springer, 2021.

[28] Adji B. Dieng, Francisco J. R. Ruiz, and David M. Blei. The dynamic embedded topic model. *CoRR*, abs/1907.05545, 2019.

[29] Aaron Schein, Scott W. Linderman, Mingyuan Zhou, David M. Blei, and Hanna M. Wallach. Poisson-randomized gamma dynamical systems. In *Advances in Neural Information Processing Systems*, pages 781–792, 2019.

[30] Dandan Guo, Bo Chen, Hao Zhang, and Mingyuan Zhou. Deep poisson gamma dynamical systems. In *Advances in Neural Information Processing Systems*, pages 8451–8461, 2018.

[31] Lantao Yu, Weinan Zhang, Jun Wang, and Yong Yu. Seqgan: Sequence generative adversarial nets with policy gradient. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2852–2858, 2017.

[32] Timothy P Lillicrap, Jonathan J Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971*, 2015.

[33] Richard S Sutton, David A McAllester, Satinder P Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. In *Advances in neural information processing systems*, pages 1057–1063, 2000.

[34] David Silver, Guy Lever, Nicolas Heess, Thomas Degris, Daan Wierstra, and Martin Riedmiller. Deterministic policy gradient algorithms. In *International conference on machine learning*, pages 387–395. PMLR, 2014.

[35] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

[36] Geoffrey Hinton, Nitish Srivastava, and Kevin Swersky. Neural networks for machine learning lecture 6a overview of mini-batch gradient descent. *Cited on*, 14(8):2, 2012.

[37] Franca Debole and Fabrizio Sebastiani. An analysis of the relative hardness of reuters-21578 subsets. *Journal of the American Society for Information Science and technology*, 56(6): 584–596, 2005.

[38] Thorsten Joachims. A probabilistic analysis of the rocchio algorithm with tfidf for text categorization. Technical report, Carnegie-mellon univ pittsburgh pa dept of computer science, 1996.

[39] David D Lewis, Yiming Yang, Tony Russell-Rose, and Fan Li. Rcv1: A new benchmark collection for text categorization research. *Journal of machine learning research*, 5(Apr): 361–397, 2004.

[40] Mark Craven, Andrew McCallum, Dan PiPasquo, Tom Mitchell, and Dayne Freitag. Learning to extract symbolic knowledge from the world wide web. Technical report, Carnegie-mellon univ pittsburgh pa school of computer Science, 1998.

[41] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[42] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

[43] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[44] Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3):229–256, 1992.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] See Abstract
   (b) Did you describe the limitations of your work? [Yes] See Appendix G
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] See supplemental material
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See Section 5.1 and Appendix D
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] See Appendix E
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See Section 5.1
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] See supplemental material
   (b) Did you mention the license of the assets? [Yes] See supplemental material
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] See supplemental material
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

Appendix

A Inference Network of dc-ETM

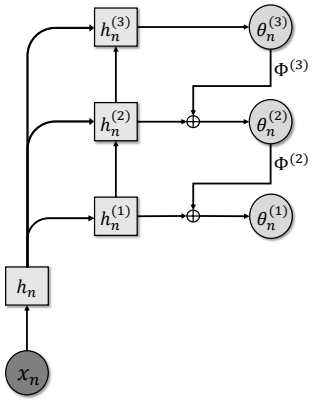


Figure 6: Overview of the hierarchical upward and downward inference network (encoder) of proposed dc-ETM.

As a usual VAE-like model, the inference network (encoder) in dc-ETM is designed to approximate the gamma distributed posteriors of latent representations of the corresponding generative model (decoder), specifically $p(\boldsymbol{\theta}_n|\boldsymbol{x}_n)$ defined in Eq. (1). As the gamma distribution, which can generate sparse and non-negative random variables, is not reparameterizable with respect to its shape parameter, we introduce the Weibull distribution [19] to construct our inference network due to its attractive properties as discussed in Appendix. B, one of which is that a latent variable $x \sim \text{Weibull}(k, \lambda)$ can be easily reparameterized as

$$x = \lambda(-\ln(1-\varepsilon))^{1/k}, \;\; \varepsilon \sim \text{Uniform}(0,1). \tag{17}$$

Specifically, following traditional VAEs, we first factorize the variational posterior distribution $q(\boldsymbol{\theta}_n|\boldsymbol{x}_n)$ defined by the inference network in a hierarchical manner as follows

$$q(\boldsymbol{\theta}_n|\boldsymbol{x}_n) = q(\boldsymbol{\theta}_n^{(L)}|\boldsymbol{x}_n) \prod_{l=1}^{L-1} q(\boldsymbol{\theta}_n^{(l)}|\boldsymbol{\theta}_n^{(l+1)}, \boldsymbol{x}_n), \tag{18}$$

which is expected to be flexible enough to well approximate the true posterior distribution $p(\boldsymbol{\theta}_n|\boldsymbol{x}_n)$. Then, for the design of the network structure, we develop a novel Weibull-based upward-downward inference network, which contains both bottom-up deterministic and top-down stochastic paths, contributing to reducing the noise in the procedure of inferencing stochastic latent variables that are higher in the hierarchy [15, 16].

As the inference network shown in the left part of 6, the bottom-up deterministic path takes the BoW vector $\boldsymbol{x}_n$ as input to obtain hierarchical deterministic latent representations $\{\boldsymbol{h}_n^{(l)}\}_{l=1}^{L}$ as follows:

$$\boldsymbol{h}_n^{(l)} = \text{MLP}(\boldsymbol{h}_n^{(l-1)} \oplus \boldsymbol{h}_n), \tag{19}$$

through specifically defining $\boldsymbol{h}_n = \text{MLP}(\boldsymbol{x}_n)$ and $\boldsymbol{h}_n^{(1)} = \text{MLP}(\boldsymbol{h}_n)$ with MLP indicating a two-layered fully connected network and $\oplus$ being the concatenation operation at feature dimension. Then, each deterministic latent representation $\boldsymbol{h}_n^{(l)}$ is further transferred into

$$\hat{\boldsymbol{k}}_n^{(l)} = \text{Relu}(\text{Linear}(\boldsymbol{h}_n^{(l)})), \tag{20}$$

$$\hat{\boldsymbol{\lambda}}_n^{(l)} = \text{Relu}(\text{Linear}(\boldsymbol{h}_n^{(l)})), \tag{21}$$

where Linear is a dense fully connected layer. Finally, taking $\boldsymbol{\Phi}^{(l+1)}\boldsymbol{\theta}_n^{(l+1)}$ as the prior information passing from the deeper layer, the variational posterior $q(\boldsymbol{\theta}_n^{(l)}|-)$ can be obtained with the stochastic

up-down path as

$$
\begin{aligned}
&q(\boldsymbol{\theta}_n^{(l)}|\boldsymbol{h}_n^{(l)}, \boldsymbol{\Phi}^{(l+1)}, \boldsymbol{\theta}_n^{(l+1)}) = \text{Weibull}(\boldsymbol{k}_n^{(l)}, \boldsymbol{\lambda}_n^{(l)}), \\
&\boldsymbol{k}_n^{(l)} = \text{Softplus}(\text{Linear}(\boldsymbol{\Phi}^{(l+1)}\boldsymbol{\theta}_n^{(l+1)} \oplus \hat{\boldsymbol{k}}_n^{(l)})), \\
&\boldsymbol{\lambda}_n^{(l)} = \text{Softplus}(\text{Linear}(\boldsymbol{\Phi}^{(l+1)}\boldsymbol{\theta}_n^{(l+1)} \oplus \hat{\boldsymbol{\lambda}}_n^{(l)})),
\end{aligned} \tag{22}
$$

where Softplus function is applied to ensure positive Weibull shape and scale parameters; $\hat{\boldsymbol{k}}_n^{(l)}$ and $\hat{\boldsymbol{\lambda}}_n^{(l)}$ incorporate the information passing from $\boldsymbol{h}_n^{(l)}$; and $\boldsymbol{\theta}_n^{(l)}$ can be obtained with the reparameterization technique defined in Eq. (17).

We emphasize that the inference network of our dc-ETM can be naturally treated as an inverse procedure of the generation process of $\boldsymbol{x}_n$, whose generation is conditional on the latent representations $\{\boldsymbol{\theta}_n^{(l)}\}_{l=1}^L$ across all hidden layers. Inverse to the generative network, the inference network can directly inject the data information into the latent representation of each layer through $\boldsymbol{h}_n$ and allow all the stochastic latent variables $\{\boldsymbol{\theta}_n^{(l)}\}_{l=1}^L$ to have a deterministic dependency on the observation $\boldsymbol{x}_n$, empirically alleviating top stochastic latent variables from being collapsed [16].

B Properties of Weibull distribution

• Similar density characteristics with Gamma Distribution

The Weibull distribution has similar characteristics with gamma distribution, *i.e.*, the density functions of the two distributions are quite similar

$$
\begin{aligned}
&\text{Weibull PDF: } P(x|k,\lambda) = \frac{k}{\lambda^k} x^{k-1} e^{(x/\lambda)^k}, \\
&\text{Gamma PDF: } P(x|\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}.
\end{aligned} \tag{23}
$$

• Easily Reparameterization

The latent variable $x \sim \text{Weibull}(k, \lambda)$ can be reparameterized as

$$
x = \lambda(-\ln(1-\varepsilon))^{1/k}, \;\; \varepsilon \sim \text{Uniform}(0,1), \tag{24}
$$

leading to an expedient and numerically stable gradient calculation.

• Analytic KL-Divergence

The KL-divergence between the Weibull and gamma distributions has an analytic expression formulated as

$$
\begin{aligned}
&\text{KL}(\text{Weibull}(k,\lambda)||\text{Gamma}(\alpha,\beta)) = -\alpha \ln \lambda + \frac{\gamma\alpha}{k} \\
&+ \ln k + \beta\lambda\Gamma(1+\frac{1}{k}) - \gamma - 1 - \alpha \ln \beta + \ln \Gamma(\alpha).
\end{aligned} \tag{25}
$$

C PG-based Training Algorithm

Modeling the generation of long sequence with RL-based methods has achieved great success, but still suffers from high variance during model training. The common ways to conduct policy gradient method for training include: **i)** employ the score-ratio gradient approximator like REINFORCE [44] to backward the gradients from Q-value $Q^\pi(s_n^{(l)}, a_n^{(l)})$ to the parameters of the policy network $\boldsymbol{\pi}$; **ii)** introduce an additional "critic" network to estimate the Q-value $Q^\pi(s_n^{(l)}, a_n^{(l)})$ as $\hat{Q}^\pi(s_n^{(l)}, a_n^{(l)})$, then backward the gradients from estimated Q-value $\hat{Q}^\pi(s_n^{(l)}, a_n^{(l)})$ to the parameters of the policy network $\boldsymbol{\pi}$ through methods like deep deterministic policy gradient (DDPG) [32]. We emphasize that, both methods above are developed for the RL applications where the Q-value function is not differentiable, e.g., the reward $r(s_n^{(l)}, a_n^{(l)})$ at every time step l is given by the non-differentiable black-box environment simulator, or for the applications with hundreds or even thousands time steps, where the direct

computation of the gradient from $Q^\pi(s_n^{(l)}, a_n^{(l)}) = r(s_n^{(l)}, a_n^{(l)}) + \mathbb{E}_\pi\left[\sum_{i=1}^{l-1}\gamma^i r(s_n^{(l-i)}, a_n^{(l-i)})\right]$ to $\boldsymbol{\pi}$ is very expensive. However, the REINFORCE estimator is notorious for its high variance, which would cause the learning of long sequence hard to converge, and the introducing of the "critic" network would also bring extra bias for the gradients and computation burden.

Moving beyond the gradient estimator of REINFORCE method or introducing the critic network, we design a more stable way to model the sequential generative process in deep topic model here. Specifically, the number of time steps here is the number of layers L of the topic model, e.g., L is 5 in this paper, which is a much smaller number compared to the hundreds of time step in other RL applications. Besides, we design the $r(s_n^{(l)}, a_n^{(l)})$ as a totally differentiable function in Eq. 14, which leads to a differentiable Q-value function $Q^\pi(s_n^{(l)}, a_n^{(l)})$. More specifically, recall the purpose of policy gradient method is to maximize the Q-value by applying gradient descent to the policy model $\boldsymbol{\pi}$, we develop the Policy Gradient based variational inference algorithm for dc-ETM in Algorithm 1.

Algorithm 1 The Policy Gradient based variational inference algorithm for dc-ETM.

```
Set minibatch size m and the number of Layer L;
Initialize the encoder parameters Ω and decoder parameters Ψ ;
for iter = 1, 2, ··· do
  Randomly select a minibatch of m documents to form a subset X = {x_j}_{1,m};
  Draw random noise {ε_i^l}_{i=1,l=1}^{m,L} from uniform distribution;
  for Layer l = L, L − 1, ··· , 1 do
    for i = l, l − 1, ··· , 1 do
      Calculate r(s_n^(i), a_n^(i)) according to Eq. (14);
    end for
    Calculate −∇_{Ω,Ψ} Q^π(s_n^(l), a_n^(l); X, {ε_i^l}_{i=1,l=1}^{m,L}) according to Eq. (15), and update Ω, Ψ
    jointly;
  end for
end for
```

D Datasets

R8 is a subset of 7,674 documents selected form 8 different review groups of the Reuters 21578 dataset, and has been split into a training set of 5,485 documents and a testing set of 2,189 ones. **20News** dataset consists of 18,774 documents from 20 various new groups and has been split into a training set of 11,314 documents and a testing set of 7,532 ones. **RCV1** is an archive of 804,114 manually categorized newswire stories made available by Reuters. **WebKB** is a dataset that includes web pages from computer science departments of various universities, which has 4,518 web pages that are categorized into 6 imbalanced categories.

E Error Bars

We randomly run 5 seeds for our method in experiments and report the error bar as below.

Table 3: Error bar for our method in the comparisons of the average of perplexities and topic diversities across all hidden layers on various benchmarks.

Model	Perplexity			Topic Diversity		
	R8	20News	RCV1	R8	20News	RCV1
dc-ETM-α	521±6	730±7	912±11	0.212±0.002	0.281±0.001	0.435±0.003
dc-ETM-β	427±5	710±5	873±10	0.346±0.003	0.429±0.003	0.566±0.005
dc-ETM-α (Policy)	463±4	707±5	896±5	0.279±0.002	0.385±0.002	0.519±0.004
dc-ETM-β (Policy)	420±3	647±4	841±5	0.379±0.004	0.456±0.003	0.584±0.004

Table 4: Error bar of our method in the document clustering comparison on the 1st hidden layer or the concatenation of all hidden layers of different topic models.

Model	Layer	WebKB		20News		R8	
		Purity	NMI	Purity	NMI	Purity	NMI
dc-ETM-α	1	61.14±0.5	26.29±0.2	32.81±0.2	43.64±0.4	75.60±0.5	39.83±0.2
	All	63.18±0.4	28.35±0.2	41.83±0.4	44.52±0.3	76.31±0.5	43.73±0.4
dc-ETM-β	1	54.71±0.5	21.43±0.1	39.80±0.2	44.30±0.3	74.30±0.6	38.63±0.3
	All	67.29±0.4	33.60±0.3	45.00±0.2	46.20±0.3	76.25±0.5	45.64±0.4
dc-ETM-α (Policy)	1	49.71±0.3	14.86±0.1	37.88±0.2	43.56±0.2	71.65±0.4	32.73±0.2
	All	64.32±0.4	33.65±0.1	42.21±0.2	45.59±0.3	77.46±0.6	44.60±0.3
dc-ETM-β (Policy)	1	57.32±0.5	26.05±0.1	40.11±0.2	44.12±0.3	71.30±0.4	38.34±0.2
	All	69.32±0.5	38.53±0.3	48.60±0.4	55.79±0.4	78.29±0.6	48.62±0.5

F Comparison of topic quality

To make an intuitive comparison on the aspect of topic quality, we visualize the 5-layer topics learned by dc-ETM and Sawtooth on 20News dataset as shown in . Obviously, the topics learned by Sawtooth are quite similar, which potentially explains the reason why concatenating its hierarchical latent document representations cannot improve and even hurt the performance on downstream tasks. On the contrary, the developed dc-ETM can learn meaningful and diverse topics in higher layers, indicating that more data information is passed to higher layers to alleviate the phenomenon of collapse.

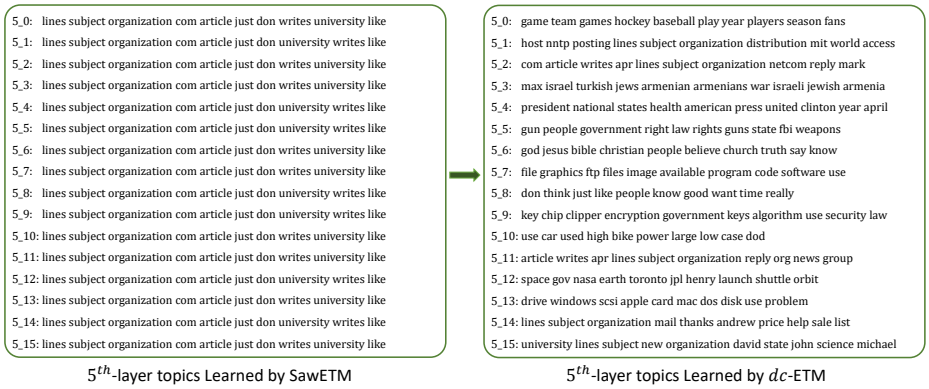


Figure 7: The 5-layer topics learned by dc-ETM and Sawtooth on 20News dataset, where each topic is interpreted by its top-10 words by sorting the word probabilities by descending order.

G Limitation

The limitation could be the computation burden brought by the extra $L-1$ training steps due to the Policy Gradient training method. However, since L is a small number, the extra computation burden is affordable. Besides, compared with the backbone GBN-based deep topic model, our method would not bring extra computation burden during the testing stage, but leads to performance improvements as shown in Section 5.

H Broader Impact

At first, we need to emphasize that this work is developed to mitigate the phenomenon of information decay in deep topic models, which has been widely disclosed in the topic modeling literature but few efforts have been made to address this challenge. The main difficulty is the need for carefully designing the probabilistic generative process to build the effective connection between the observation and the corresponding latent representations, on the premise of preserving the interpretable hierarchical topic modeling structure, rather than casually introducing skip-connections. Utilizing the natural hierarchy in deep topic models, we provide a general solution equipped with a novel perspective to incorporate RL-based training algorithm, which brings quite a few contributing ideas to this field.

Please note that, we take an NTM named SawETM as an example in this paper to illustrate how we develop a dc-ETM, but dc-ETM can be also applied to extend NTMs with similar structures. Most of exsiting NTMs adopt VAE-liked structures, but none of them attempt to solve the "*posterior collapse*" phenomenon in essence, resulting in that the latent representations at higher layers exhibit similar or meaningless patterns (as shown in the Appendix F).

The significance of developing deep-coupling structure with RL-based training algorithm, which has effectively improved the quality of the latent representations of a deep topic model at higher layers, goes beyond a single specific model.

I The implementation details of dc-ETM variants

To have an intuitive understanding about the implementation details of the projection metrics in dc-ETM, we provide an overview of the network structure of a 3-layer dc-ETM in Fig. 8. As shown in Fig. 8, the topic matrix $\boldsymbol{\Phi}^{(l)}$ at layer l can be factorized into the product of two topic embedding matrices. Then, there are two kinds of choices to obtain the projection matrices, leading to the variants distinguished by the suffix $-\alpha$ and $-\beta$.

Specifically, for the variant dc-ETM-α defined in Eq. (2), the projection matrix $\hat{\boldsymbol{\Phi}}^{(3)}$ can be obtained by multiplying $\boldsymbol{\Phi}^{(1)}\boldsymbol{\Phi}^{(2)}\boldsymbol{\Phi}^{(3)}$, and then the augmented observation vector $\boldsymbol{x}_n^{(3)}$ can be generated from the Poisson distribution with a rate of $\alpha^{(3)}\hat{\boldsymbol{\Phi}}^{(3)}\boldsymbol{\theta}_n^{(3)}$. The other projection matrices $\hat{\boldsymbol{\Phi}}^{(2)}$ and $\hat{\boldsymbol{\Phi}}^{(1)}$ can be obtained in a simliar way, where specifically defining $\hat{\boldsymbol{\Phi}}^{(1)} := \boldsymbol{\Phi}^{(1)}$.

For the variant dc-ETM-β defined in Eq. (3), the projection matrix $\hat{\boldsymbol{\Phi}}^{(3)}$ can be obtained by directly multiplying $\boldsymbol{\beta}^{(0)T}$ and $\boldsymbol{\beta}^{(3)}$, which is more efficient than dc-ETM-α. Thus, in our consideration, with a shorter path for gradient propagation, the short connections in dc-ETM-β can perverse more data information than those in dc-ETM-α, resulting in more informative latent document representations to achieve better model performance with less "*posterior collapse*".

We emphasize that, the projection in dc-ETM-α could be applied to extend any existing deep topic models, while the projection in dc-ETM-β is only applicable for the NTMs equipped with topic embedding techniques.

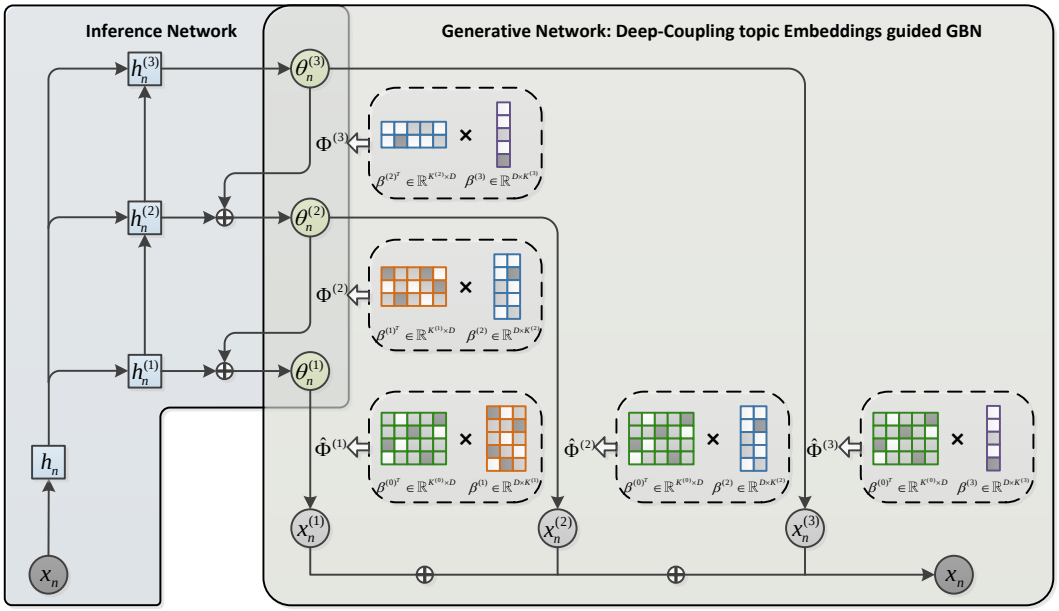


Figure 8: Overview of the proposed dc-ETM, where the left part is the hierarchical upward and downward inference network (encoder) and the right part is the generative network (decoder).

J Preliminary of "*posterior collapse*"

J.1 Definition of "*posterior collapse*"

For a VAE-based model consisting of a decoder $p_\theta(\boldsymbol{x}|\boldsymbol{z})$ and an encoder $q_\phi(\boldsymbol{z}|\boldsymbol{x})$, the definition of posterior collapse is that the posterior of latent variables, denoted as $q_\phi(\boldsymbol{z}|\boldsymbol{x})$, collapses to it prior $p_\theta(\boldsymbol{z})$, which is a non-informative distribution and independent of the data $\boldsymbol{x}$. It can be mathematically denoted as that the KL divergence between $q_\phi(\boldsymbol{z}|\boldsymbol{x})$ and $p_\theta(\boldsymbol{z})$ is close to zero, represented as $D_{KL}[q_\phi(\boldsymbol{z}|\boldsymbol{x})||p(\boldsymbol{z})] \approx 0$.

J.2 How to measure "*posterior collapse*"

As the definition of "*posterior collapse*", a promising metric to measure "*posterior collapse*" could be the KL-divergence between $q_\phi(\boldsymbol{z}|\boldsymbol{x})$ and $p_\theta(\boldsymbol{z})$, where a smaller KL-divergence score indicates that the posterior $q_\phi(\boldsymbol{z}|\boldsymbol{x})$ contains less data information and has a larger tendency to collapse to its non-informative prior. In Table 5, we compare our method dc-ETM-β(policy) (for brevity, we note it as "dc-ETM") with the previous SOTA NTM named SawETM with the metric of layer-wise KL divergence on 20News, R8, and RCV1 datasets. As the results shows, the KL divergence of SawETM gradually reduces to zero with the network going deeper, which indicates the occurrence of a serious degree of "*posterior collapse*" at higher layers and limited data information can be captured by these latent variables at higher layers. On the contrary, benefiting from the efficient skip connection between the observation $\boldsymbol{x}_n$ and multiple latent document representations $\{\boldsymbol{\theta}_n^{(l)}\}_{l=1}^L$, dc-ETM can achieve relatively larger KL-divergence scores at higher layers by alleviating "*posterior collapse*", leading to more expressive latent document representations for downstream tasks.

KL-divergence of each layer						
Metric	20News		R8		RCV1	
Layer #i	SawETM	dc-ETM	SawETM	dc-ETM	SawETM	dc-ETM
1	124	354	80.5	161	156	365
2	38.6	233	26.1	89.8	72.1	235
3	3.07	158	11.7	81.9	55.4	170
4	2.51	132	2.05	67.4	48.7	122
5	1.64	101	1.16	77.7	48.0	102

Table 5: Layer-wise KL divergence scores of SawETM and dc-ETM.

J.3 Why "*posterior collapse*" in a VAE-based model

Besides the experimental analysis, to have a theoretical understanding of the reason why *"posterior collapse"* widely exists in the higher layers of hierarchical VAEs, we try to explain the phenomenon of *"posterior collapse"* from the perspective of information theory. For ease of understanding, we use a vanilla L-layer hierarchical VAE with a top-down inference network as an example, where dc-ETMs can be all treated as VAE-based models. Specifically, following [17], we extend their theoretical explanation for *"posterior collapse"* in a single-layer VAE to a multi-layer version.

Separating the latent variables as the lower-level variables $\boldsymbol{z}_{\leq k} = \{\boldsymbol{z}_1, ..., \boldsymbol{z}_k\}$ and the higher-level ones $\boldsymbol{z}_{>k} = \{\boldsymbol{z}_{k+1}, ..., \boldsymbol{z}_L\}$, where $k \in \{0, ..., L-1\}$, then the ELBO for this hierarchical VAE can be reformulated as

$$\mathcal{L} = \mathbb{E}_{p(\boldsymbol{x})}\left[\mathbb{E}_{q_\phi(\boldsymbol{z}_{\leq k}|\boldsymbol{z}_{>k})q_\phi(\boldsymbol{z}_{>k}|\boldsymbol{x})}\left[\log p_\theta(\boldsymbol{x}|\boldsymbol{z}_1)\right] - \sum\nolimits_{l=1}^{L} D_{\mathrm{KL}}(q_\phi(\boldsymbol{z}_l|\boldsymbol{z}_{l+1})||p_\theta(\boldsymbol{z}_l|\boldsymbol{z}_{l+1}))\right], \tag{26}$$

where $q_\phi(\boldsymbol{z}_L|\boldsymbol{z}_{L+1}) := q_\phi(\boldsymbol{z}_L|\boldsymbol{x})$, $p_\theta(\boldsymbol{z}_L|\boldsymbol{z}_{L+1}) := p_\theta(\boldsymbol{z}_L)$, and the main contribution to the expected log-likelihood term is coming from the lower-level latent variables $\boldsymbol{z}_{\leq k}$ before the k-th hidden layer [16]. Once the generation capacity of the generative model $p_\theta(\boldsymbol{x}|\boldsymbol{z}_{\leq k})$ is powerful enough to reconstruct the observation $\boldsymbol{x}$ well, the variational posteriors of higher-level latent variables $\boldsymbol{z}_{>k}$ will be optimized to be close to their priors, i.e., $q_\phi(\boldsymbol{z}_{>k}|\boldsymbol{x}) \approx p_\theta(\boldsymbol{z}_{>k})$, leading the representations learned by VAE at higher layers to be meaningless and cannot provide faithful summaries for $\boldsymbol{x}$, which is well-known as the phenomenon of "*posterior collapse*" or "*latent variable collapse*" [17].

To find the potential solutions to alleviating "*posterior collapse*", in the following, we reinterpret this phenomenon from the perspective of information theory by extending the findings in [17] to a

hierarchical VAE scenario. For ease of undertanding, we define the mutual information between the data $\boldsymbol{x}$ and the higher-level latent variables $\boldsymbol{z}_{>k}$ as

$$\mathcal{I}_q(\boldsymbol{x}, \boldsymbol{z}_{>k}) = -\mathcal{H}_q(\boldsymbol{z}_{>k}|\boldsymbol{x}) + \mathcal{H}_q(\boldsymbol{z}_{>k}) = \mathbb{E}_{p(\boldsymbol{x})q_\phi(\boldsymbol{z}_{>k}|\boldsymbol{x})} \log q_\phi(\boldsymbol{z}_{>k}|\boldsymbol{x}) - \mathbb{E}_{q_\phi(\boldsymbol{z}_{>k})} \log q_\phi(\boldsymbol{z}_{>k}),$$

which is induced by the variational posterior $q_\phi(\boldsymbol{z}_{>k}|\boldsymbol{x})$. Then KL term in Eq. (26) can be rewritten as

$$\begin{aligned}
&\mathbb{E}_{p(\boldsymbol{x})}\left[\sum\nolimits_{l=1}^{L} D_{\mathrm{KL}}(q_\phi(\boldsymbol{z}_l|\boldsymbol{z}_{l+1})||p_\theta(\boldsymbol{z}_l|\boldsymbol{z}_{l+1}))\right] \\
=&\mathbb{E}_{p(\boldsymbol{x})}\left[\sum\nolimits_{l=1}^{k} D_{\mathrm{KL}}(q_\phi(\boldsymbol{z}_l|\boldsymbol{z}_{l+1})||p_\theta(\boldsymbol{z}_l|\boldsymbol{z}_{l+1}))\right] + \mathbb{E}_{p(\boldsymbol{x})}\left[D_{\mathrm{KL}}(q_\phi(\boldsymbol{z}_{>k}|\boldsymbol{x})||p_\theta(\boldsymbol{z}_{>k}))\right] \\
=&\mathbb{E}_{p(\boldsymbol{x})}\left[\sum\nolimits_{l=1}^{k} D_{\mathrm{KL}}(q_\phi(\boldsymbol{z}_l|\boldsymbol{z}_{l+1})||p_\theta(\boldsymbol{z}_l|\boldsymbol{z}_{l+1}))\right] + \mathcal{I}_q(\boldsymbol{x}, \boldsymbol{z}_{>k}) + D_{\mathrm{KL}}(q_\phi(\boldsymbol{z}_{>k})||p_\theta(\boldsymbol{z}_{>k})),
\end{aligned} \tag{27}$$

where $q_\phi(\boldsymbol{z}_{>k}) = \mathbb{E}_{p(\boldsymbol{x})}\left[q_\phi(\boldsymbol{z}_{>k}|\boldsymbol{x})\right]$. By substituting Eq. (27) into Eq. (26), due to the non-negativity of mutual information and KL divergence, we can find that maximizing the ELBO is opposite to maximizing the mutual information $\mathcal{I}_q(\boldsymbol{x}, \boldsymbol{z}_{>k})$. When $\mathcal{I}_q(\boldsymbol{x}, \boldsymbol{z}_{>k})$ is minimized to zero, the variational posterior $q_\phi(\boldsymbol{z}_{>k}|\boldsymbol{x})$ will be independent of the data $\boldsymbol{x}$, which leads to the phenomenon of "*posterior collapse*".

Thus, exploiting the property of the deep NTMs and building the skip connection between the $\boldsymbol{x}_n$ and $\boldsymbol{\theta}_n^{(l)}$ with two projection variants, as shown in Fig. 8, is promising to force the higher layers of latent variables be more informative by maximizing the mutual information.

K Evaluations of topic quality under more metrics

For evaluating the quality of the learned topics, we introduce more metrics for comparison between SawETM and dc-ETM-β (policy) (for brevity, we note it as "dc-ETM") including: topic diversity, topic coherence, and topic quality (a product of the topic diversity and topic coherence). As shown in Table 6, Table 7, and Table 8, dc-ETM can achieve comparable performance on the aspect of topic quality at the first layer while significantly outperform SawETM in the higher layers.

R8						
Metric	Diversity		Coherence		Topic Quality	
Layer #i	SawETM	dc-ETM	SawETM	dc-ETM	SawETM	dc-ETM
1	42.11	46.56	43.29	48.44	0.182	0.225
2	15.50	37.38	44.68	41.29	0.069	0.154
3	13.75	36.48	45.05	40.18	0.062	0.146
4	12.34	34.09	47.78	47.25	0.059	0.161
5	20.00	35.43	40.89	50.79	0.082	0.179
Average	20.70	37.98	44.34	45.59	0.091	0.173

Table 6: Topic Diversity, Coherence, and Quality (product of diversity and coherence) for each layer of SawETM and dc-ETM on R8 dataset.

20News						
Metric	Diversity		Coherence		Topic Quality	
Layer #i	SawETM	dc-ETM	SawETM	dc-ETM	SawETM	dc-ETM
1	38.82	33.05	31.01	32.32	0.120	0.107
2	22.53	40.04	55.91	54.69	0.126	0.218
3	9.531	53.75	63.67	56.21	0.067	0.302
4	9.310	47.96	59.42	60.77	0.055	0.291
5	7.309	53.02	62.04	63.23	0.045	0.335
Average	17.50	45.56	54.51	53.44	0.083	0.251

Table 7: Topic Diversity, Coherence, and Quality (product of diversity and coherence) for each layer of SawETM and dc-ETM on 20News dataset.

RCV1						
Metric	Diversity		Coherence		Topic Quality	
Layer #i	SawETM	dc-ETM	SawETM	dc-ETM	SawETM	dc-ETM
1	59.74	66.01	38.30	36.04	0.229	0.237
2	49.45	53.01	39.58	39.66	0.196	0.210
3	20.01	57.03	34.08	44.17	0.068	0.251
4	14.84	55.15	31.27	45.79	0.046	0.253
5	21.87	60.93	30.94	48.55	0.067	0.295
Average	33.18	58.42	34.83	42.84	0.121	0.249

Table 8: Topic Diversity, Coherence, and Quality (product of diversity and coherence) for each layer of SawETM and dc-ETM on R8 dataset.

L Layer-wise comparison under the clustering task

We provide an additional evaluation for the clustering tasks on each layer of SawETM and dc-ETM-β(policy) (for brevity, we note it as "dc-ETM"). As shown in Table 9, the higher layer $\boldsymbol{\theta}_n^{(l)}$ of SawETM can only preserve limited data information for the downstream document clustering task, which could explain why the SawETM cannot obtain a gain of performance improvement even if concatenating all hidden layers, specifically $\{\boldsymbol{\theta}_n^{(l)}\}_{l=1}^{L}$ (denoted as "All").

Datasets	20News				R8			
Metric	Purity		NMI		Purity		NMI	
Layer #i	SawETM	dc-ETM	SawETM	dc-ETM	SawETM	dc-ETM	SawETM	dc-ETM
1	43.33	40.11	50.77	44.12	75.25	71.30	42.97	38.34
2	33.52	42.22	45.88	46.72	76.01	75.38	38.41	40.14
3	26.72	44.50	40.12	47.03	34.38	74.22	27.09	44.75
4	12.51	45.76	30.20	50.26	22.10	75.00	17.03	43.10
5	11.82	46.89	26.78	53.81	20.17	76.19	15.20	45.89
All	38.69	48.60	39.33	55.79	75.89	78.29	39.55	48.62

Table 9: Comparison of different layers' $\boldsymbol{\theta}_n^{(l)}$ quality under the clustering task for SawETM and dc-ETM.