Bayesian Deep Embedding Topic Meta-Learner

Zhibin Duan [*1] **Yishi Xu** [*1] **Jianqiao Sun** [1] **Bo Chen** [1] **Wenchao Chen** [1] **Chaojie Wang** [1] **Mingyuan Zhou** [2]

Abstract

Existing deep topic models are effective in capturing the latent semantic structures in textual data but usually rely on a plethora of documents. This is less than satisfactory in practical applications when only a limited amount of data is available. In this paper, we propose a novel framework that efficiently solves the problem of topic modeling under the small data regime. Specifically, the framework involves two innovations: a bi-level generative model that aims to exploit the task information to guide the document generation, and a topic meta-learner that strives to learn a group of global topic embeddings so that fast adaptation to the task-specific topic embeddings can be achieved with a few examples. We apply the proposed framework to a hierarchical embedded topic model and achieve better performance than various baseline models on diverse experiments, including few-shot topic discovery and few-shot document classification.

1. Introduction

Topic models (Blei et al., 2003) enjoy great popularity among various text mining tools. Their prevalence stems from their ability to organize a collection of documents into a set of prominent themes. In addition to highlighting the underlying patterns intuitively, these extracted topics can also be used to derive low-dimensional representations of the documents, which have proven useful in a series of natural language processing tasks, such as information retrieval (Wang et al., 2007), text classification (Rubin et al., 2012), and machine translation (Mimno et al., 2009).

Over the recent years, considerable progress has been made in topic modeling, ranging from exploring hierarchical document representations (Zhou et al., 2016; Guo et al., 2018) to discovering more coherent topics (Bianchi et al., 2021). In particular, the development of variational autoencoder (VAE) (Kingma & Welling, 2013) shows the potential of deep neural networks in posterior inference, motivating the proposal of an array of neural topic models (NTMs) (Miao et al., 2016; Zhang et al., 2018; Dieng et al., 2020) that possess better flexibility and scalability. However, relatively little work has focused on few-shot adaptation in topic models (Iwata, 2021). This is partly because a topic model is often supposed to deal with a large volume of text, but more primarily due to the challenging nature of learning joint distributions from a few samples in an unsupervised manner. On the other side, there are inevitably times in real-world applications when only a limited amount of data is available, such as written records of ancient history or news from front-line reporters. Thus, it is crucial to explore effective ways for few-shot learning in topic models.

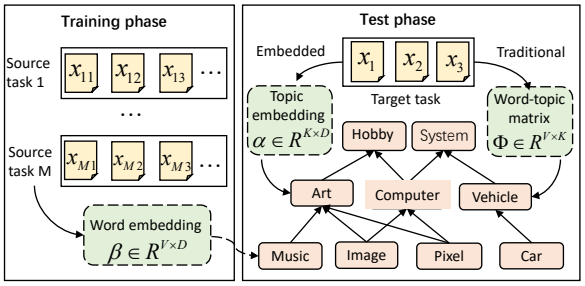


Figure 1. Illustration of the advantage of embedded topic models over traditional topic models in the few-shot setting. Considering word embeddings as transferable knowledge reduces the difficulty of discovering new topics from novel target tasks.

Few-shot learning is a long-standing problem that aims to efficiently solve new tasks with only a few examples by exploiting knowledge gained from a large number of related tasks (Kim et al., 2019). While there have been many successful learning paradigms, most of them were originally designed for supervised tasks and are not natural to be applied in topic models. Basically, the form of knowledge that can be transferred to new tasks plays a key role in determining a few-shot algorithm. For instance, the model-agnostic meta-learning (MAML) (Finn et al., 2017) algorithm defines the knowledge as a good initialization of model parameters that can be adapted with a few gradient descent steps to

[*]Equal contribution [1]National Laboratory of Radar Signal Processing, Xidian University, Xi'an, China. [2]McCombs School of Business, The University of Texas at Austin, Austin, TX 78712, USA. Correspondence to: Bo Chen <bchen@mail.xidian.edu.cn>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022.

individual tasks. Similarly, the knowledge in topic modeling could be designed as a meaningful prior of the topics. However, compared to traditional topic models that adapt the entire topics to each task, we find that the embedded topic model (ETMs) (Dieng et al., 2020; Duan et al., 2021a) can be more effective in achieving the few-shot adaptation. Concretely, as shown in Fig. 1, the word embeddings serve as transferable knowledge that can be learned from multiple training tasks, as a result, only the topic embeddings need to be inferred when adapting to a new task.

Although the embedded topic model somewhat alleviates the plight of discovering new topics from a few documents, it does not essentially solve the problem of topic modeling under the small data regime. To this end, we draw inspiration from several recent studies that cast meta-learning as Bayesian inference in hierarchical modelling (Gordon et al., 2018; Ravi & Beatson, 2018) and further propose to learn a group of global topic embeddings, which are modeled as Gaussian distributions, providing a good quantification of uncertainty and also facilitating the fast adaptation to task-specific topic embeddings. Moreover, considering the dependence between documents and the task they belong to, we develop a bi-level generative model that captures the topic proportion information from both tasks and documents. Specifically, the task-level latent variables act as a prior to be conditioned on the document-level latent variables, which can better reflect the intrinsic variability within the tasks. In conclusion, we in this paper propose a novel framework that effectively solves the problem of learning a topic model from a few documents. We apply this framework to a hierarchical embedded topic model (Duan et al., 2021a) and achieve better performance than various baseline models in few-shot experiments. Our main contributions are summarized as follows:

- We propose a novel framework to solve the problem of topic modeling under the few-shot settings, including three novel modules: i) the word embeddings sharing mechanism of ETMs is used to reuse knowledge from training tasks; ii) a group of global topics is employed to facilitate the fast adaptation to task-specific topic embeddings; iii) A bi-level generative model is developed to capture the topic proportion information from both tasks and documents.
- The proposed framework is applied in a hierarchical topic model to obtain a novel bi-level hierarchical generative model. Meanwhile, we design an efficient upward-downward inference network to approximate the posterior for its latent variables.
- A flexible training algorithm with an episodic training strategy is developed to train the model parameters. Experiments on various datasets and tasks show that our models outperform various baseline models on few-shot learning settings.

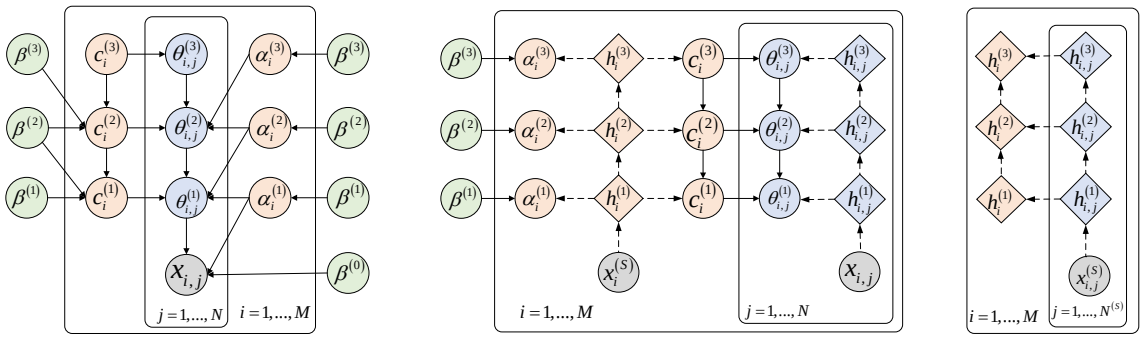
2. Related work

2.1. Topic modeling

The family of topic models has been significantly expanded in recent years, resulting in hierarchical topic models (Blei et al., 2010; Paisley et al., 2014; Gan et al., 2015; Henao et al., 2015; Zhou et al., 2015; Cong et al., 2017; Zhao et al., 2018), neural topic models (Miao et al., 2016; Srivastava & Sutton, 2017; Card et al., 2017; Zhang et al., 2018), embedded topic models (Dieng et al., 2020; Duan et al., 2021a), optimal transport-based topic models (Huynh et al., 2020; Zhao et al., 2021; Wang et al., 2022), and knowledge-guided topic models (Duan et al., 2021b). Besides, topic models can also be improved by drawing upon experiences from other domains, such as transfer learning (Hu et al., 2015) and continual lifelong learning (Gupta et al., 2020). However, little attention has been paid to the adaptability of topic models in few-shot settings. Iwata (2021) develops a few-shot learning algorithm based on latent Dirichlet allocation, the core idea of which is to learn good model priors via a shared inference network. But it is limited by the shallow structure and only explores the single-layer semantic information. Moreover, it relies on an iterative procedure to adapt to a new task at the testing stage, which is less than satisfactory when real-time processing is desired (Zhang et al., 2018). The most significant difference between our model and the above topic models is that we have introduced a task-level latent variable and maintained a group of topic embeddings shared by different tasks. They act as a good prior, facilitating accurate posterior approximations even with only a few documents.

2.2. Meta-learning

Meta-learning, also known as learning to learn, comprises a broad family of techniques focused on helping deep models quickly adapt to new environments. Existing literature commonly categorizes meta-learning approaches into three groups: i) the metric-based, ii) the model-based, and iii) the optimization-based. Metric-based methods (Vinyals et al., 2016; Snell et al., 2017; Sung et al., 2018) aim to learn a good metric space in which efficient predictions can be made for new tasks. This type of method is typically used in supervised learning. Model-based methods (Edwards & Storkey, 2016; Santoro et al., 2016; Mishra et al., 2017) depend on a well-designed module to maintain an internal state for each task. Optimization-based methods (Ravi & Larochelle, 2016; Finn et al., 2017) involve directly optimizing a meta learner that can efficiently update its parameters using a small amount of data. Some recent studies (Gordon et al., 2018; Ravi & Beatson, 2018; Iakovleva et al., 2020; Jeon et al., 2022) also cast meta-learning as Bayesian inference in a hierarchical graphical model. This approach provides a principled framework to reason about uncertainty.



(a) Generative model (b) Inference network (c) Invariant net

Figure 2. Graphical illustrations: (a) The bi-level hierarchical generative model is composed by two collections of hierarchical latent variables, $\boldsymbol{c}_i^{(l)}$ for the task $\mathcal{D}_i$ and $\theta_{i,j}^{(l)}$ for the samples $x_{i,j}$; the embedding-based topic meta-learner is a hierarchical generative model, where task-specific topic embeddings $\alpha_i^{(l)}$ are generated from global topic embeddings $\beta^{(l)}$. (b) The proposed hierarchical upward and downward encoder network for inferring document-specific latent variables $\theta_{i,j}^{(l)}$ and task-specific latent variables $\{c_i^{(l)}, \alpha_i^{(l)}\}$; (c) The proposed hierarchical permutation invariant neural network for getting hierarchical task representation $h_i^{(l)}$.

Our work can be viewed as a combination of the model-based methods and Bayesian approaches, because we use a model-based network to infer the task and document representations and a Bayesian decoder to quickly adapt to the new topic embeddings.

3. Proposed model

3.1. Problem formulation

Before diving into the details of our method, we first introduce the problem definition of few-shot learning for topic modeling. Given the bag-of-words (BoWs) representations of C training corpora $\{\boldsymbol{X}_c\}_{c=1}^{C}$, the goal of few-shot learning is to train a topic model on these corpora so that during the evaluation stage, the trained model can be adapted to discover new topics with only a few documents from a new corpus $\boldsymbol{X}_{\text{new}}$. Following the convention in most few-shot learning approaches, we adopt an episodic training strategy that samples a batch of tasks $\{\mathcal{D}_i\}_{i=1}^{M}$ from the training corpora. Each task consists of a small training subset $\mathcal{D}_i^{(S)}$ called *support set* and a validation subset $\mathcal{D}_i^{(Q)}$ called *query set*, $\mathcal{D}_i^{(S)} \cap \mathcal{D}_i^{(Q)} = \varnothing$. During evaluation, we similarly sample the test task $\mathcal{D}_{\text{test}}$ from the new corpus $\boldsymbol{X}_{\text{new}}$, use its support set $\mathcal{D}_{\text{test}}^{(S)}$ to adapt the model, and use its query set $\mathcal{D}_{\text{test}}^{(Q)}$ to evaluate the performance of the adapted model.

3.2. Bi-level hierarchical generative model

In this section, we combine the proposed framework with the gamma belief network (Zhou et al., 2015), resulting in a bi-level hierarchical topic model. Formally, the generative model with L latent layers can be expressed as

$$
\begin{aligned}
\boldsymbol{c}_i^{(L)} &\sim \text{Gam}\left(\boldsymbol{r}, e_i^{(L+1)}\right), \cdots, \\
\boldsymbol{c}_i^{(l)} &\sim \text{Gam}\left(\boldsymbol{\psi}^{(l+1)}\boldsymbol{c}_i^{(l+1)}, e_i^{(l+1)}\right), \cdots, \\
\boldsymbol{\theta}_{i,j}^{(L)} &\sim \text{Gam}\left(\boldsymbol{c}_i^{(L)}, e_j^{(L+1)}\right), \cdots, \\
\boldsymbol{\theta}_{i,j}^{(l)} &\sim \text{Gam}\left(\boldsymbol{\Phi}_i^{(l+1)}\boldsymbol{\theta}_{i,j}^{(l+1)} + \boldsymbol{c}_i^{(l)}, e_j^{(l+1)}\right), \cdots, \\
\boldsymbol{\theta}_{i,j}^{(1)} &\sim \text{Gam}\left(\boldsymbol{\Phi}_i^{(2)}\boldsymbol{\theta}_{i,j}^{(2)} + \boldsymbol{c}_i^{(1)}, e_j^{(2)}\right), \\
\boldsymbol{x}_{i,j} &\sim \text{Pois}\left(\boldsymbol{\Phi}_i^{(1)}\boldsymbol{\theta}_{i,j}^{(1)}\right),
\end{aligned}
\tag{1}
$$

where $\boldsymbol{x}_{i,j} \in \mathbb{Z}^V$ denotes the word count vector of the j^{th} document in the i^{th} task, which is factorized as the product of a task-specific factor loading matrix $\boldsymbol{\Phi}_i^{(1)} \in \mathbb{R}_+^{V \times K_1}$ and a document-specific factor score $\boldsymbol{\theta}_{i,j}^{(1)} \in \mathbb{R}_+^{K_1}$ under the Poisson likelihood. Then, in order to obtain hierarchical document representations, the shape parameter of the factor score $\boldsymbol{\theta}_{i,j}^{(l)} \in \mathbb{R}_+^{K_l}$ at layer l is further decomposed into the sum of $\boldsymbol{\Phi}_i^{(l+1)}\boldsymbol{\theta}_{i,j}^{(l+1)}$ and $\boldsymbol{c}_i^{(l)}$, where $\boldsymbol{c}_i^{(l)} \in \mathbb{R}_+^{K_l}$ is the task-specific factor score that serves as a prior incorporated with task information, till the top layer, $\boldsymbol{\theta}_{i,j}^{(L)}$ only depends on $\boldsymbol{c}_i^{(L)}$. Moreover, to build the dependence between task-specific factor scores of different layers, the shape parameter of the factor score $\boldsymbol{c}_i^{(l)}$ is factorized into $\boldsymbol{\psi}^{(l+1)}\boldsymbol{c}_i^{(l+1)}$. In addition, the set of variables in Eq. (1) $\boldsymbol{r}, e_i^{(L+1)}, \cdots, e_i^{(2)}, e_j^{(L+1)}, \cdots, e_j^{(2)}$ are all hyperparameters, which are set to fixed values following the previous work (Zhang et al., 2018), *i.e.*, $\{\boldsymbol{r} = \boldsymbol{1}, e_i^{(l)} = 1, e_j^{(l)} = 1\}$.

Distinct from other few-shot generative models (Oord et al., 2017; Giannone & Winther, 2021) that typically employ a shared decoder for all tasks, here we use a task-specific

decoder $\mathbf{\Phi}_i^{(l)}$ on account of the unique transferability of text representations. More concretely, most existing few-shot generative models mainly focus on vision tasks. As low-level patterns ($e.g.$, points and edges) in images are ubiquitous in almost every task, their corresponding representations can be shared across tasks. The situation is quiet different for textual data, where our tasks operate at the lexical level and words that frequently occur in one task may disappear in other tasks. Consequently, a task-specific decoder is of vital importance to capture the variability between different tasks.

3.3. Embedding-based topic meta-learner

As discussed in Sec. 1, embedded topic models lay a solid foundation for achieving the few-shot adaptation to novel tasks. However, unlike the vanilla embedded topic model (Dieng et al., 2020) where the unique word-topic matrix is directly formulated as the product of word embeddings and topic embeddings, the challenge of our generative model lies in that the hierarchical factor loading matrices are assumed to be independent. Fortunately, a recently proposed neural topic model called SawETM (Duan et al., 2021a) addressed this issue by putting forward a Sawtooth Connection (SC) module. Particularly, representing topics of all layers into a shared embedding space enables the factor loading matrix $\mathbf{\Phi}_{i,k}^{(l)}$ to be calculated based on the topic embeddings at two adjacent layers. As a result, the dependency between factor loading matrices of two adjacent layers is built by reusing the topic embeddings of intermediate layers.

Building on top of SawETM, we intend to further improve our model's ability to quickly learn a set of new topics from a few target documents. To this end, we resort to the concept of Bayesian meta-learning. By imposing probabilistic distributions on the network weights, it aims to learn reasonable priors that can produce accurate posterior approximations with a few steps of gradient descent or Bayesian updating (Ravi & Beatson, 2018). Nevertheless, we emphasize that it is difficult to fine-tune all parameters effectively using only a few documents, especially for the high-dimensional word embeddings. Accordingly, we design a solution where only topic embeddings need to be inferred adaptively from the new tasks. Specifically, as shown in Fig. 2(a), we maintain a group of global topic embeddings dedicated to providing good priors for task-specific topic embeddings. Thus, the generation of factor loading matrices is expressed as

$$
\begin{aligned}
&\boldsymbol{\beta}^{(l)} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}), \quad l = 0, \cdots, L \\
&\boldsymbol{\psi}_k^{(l)} = \text{softmax}\left(\boldsymbol{\beta}^{(l-1)^T} \boldsymbol{\beta}_k^{(l)}\right), \quad l = 2, \cdots, L \\
&\boldsymbol{\alpha}_i^{(l)} \sim \mathcal{N}\left(\boldsymbol{\beta}^{(l)}, \boldsymbol{I}\right), \quad l = 1, \cdots, L \\
&\mathbf{\Phi}_{i,k}^{(l)} = \text{softmax}\left(\boldsymbol{\alpha}_i^{(l-1)^T} \boldsymbol{\alpha}_{i,k}^{(l)}\right), \quad l = 1, \cdots, L
\end{aligned}
\tag{2}
$$

where $\boldsymbol{\beta}^{(l)} \in \mathbb{R}^{K_l \times D}$ represents the set of global topic embeddings shared across different tasks and $\boldsymbol{\alpha}_i^{(l)} \in \mathbb{R}^{K_l \times D}$ denotes the group of task-specific topic embeddings that is sampled from a Gaussian distribution with $\boldsymbol{\beta}^{(l)}$ as its mean vector. Note that in the bottom layer $\boldsymbol{\beta}^{(0)} \in \mathbb{R}^{V \times D}$ is the word embeddings, which is learned from a large number of training tasks and no longer fine-tuned for the target task.

3.4. Approximate inference

To perform efficient inference for the latent variables of our proposed model, we develop an effective inference network to approximate the posteriors of $\{\boldsymbol{\theta}_{i,j}^{(l)}, \boldsymbol{c}_i^{(l)}, \boldsymbol{\beta}^{(l)}, \boldsymbol{\alpha}_i^{(l)}\}_{l=1}^L$ based on amortized variational inference (AVI) techniques (Hoffman et al., 2013; Kingma & Welling, 2013).

Document latent variable inference Instead of using Gaussian latent variables as in most neural topic models (Srivastava & Sutton, 2017), our generative model employs gamma-distributed latent variables that are more suitable for modeling sparse and non-negative document representations. However, such a choice also brings the difficulty of reparameterizing gamma-distributed random variables when we design a sampling-based inference network. To mitigate this issue, we utilize a Weibull upward-downward variational encoder to approximate the posteriors of $\{\boldsymbol{\theta}_{i,j}^{(l)}\}_{l=1}^L$ inspired by the work in Zhang et al. (2018; 2020). Then we have

$$
q(\boldsymbol{\theta}_{i,j}^{(l)} \,|\, \boldsymbol{x}_{i,j}, \boldsymbol{\theta}_{i,j}^{(l+1)}, \boldsymbol{c}_j^{(l)}) = \text{Weibull}\left(\boldsymbol{k}_{i,j}^{(l)}, \boldsymbol{\lambda}_{i,j}^{(l)}\right), \tag{3}
$$

where parameters $\boldsymbol{k}_{i,j}^{(l)}, \boldsymbol{\lambda}_{i,j}^{(l)} \in \mathbb{R}_+^{K_l}$ are deterministic transformations of the observed document features, the information from the stochastic task path $\boldsymbol{c}_i^{(l)}$, and the information from the stochastic up-down path $\boldsymbol{\theta}_{i,j}^{(l+1)}$. Fig. 2(b) shows how these pieces of information are propagated to influence $\boldsymbol{\theta}_{i,j}^{(l+1)}$. Formally, the inference process can be described by

$$
\begin{aligned}
&\boldsymbol{h}_{i,j}^{(l)} = \text{Re}LU(\boldsymbol{h}_{i,j}^{(l-1)} \boldsymbol{W}_1^{(l)} + \boldsymbol{b}_1^{(l)}), \\
&\boldsymbol{h}_{i,j}^{(l)'} = \begin{cases} \boldsymbol{h}_{i,j}^{(L)} \oplus \boldsymbol{c}_i^{(L)} & l = L, \\ \boldsymbol{h}_{i,j}^{(l)} \oplus \mathbf{\Phi}_i^{(l+1)} \boldsymbol{\theta}_{i,j}^{(l+1)} \oplus \boldsymbol{c}_i^{(l)} & l < L, \end{cases} \\
&\boldsymbol{k}_{i,j}^{(l)} = \ln[1 + \exp(\boldsymbol{h}_{i,j}^{(l)'} \boldsymbol{W}_2^{(l)} + \boldsymbol{b}_2^{(l)})], \\
&\boldsymbol{\lambda}_{i,j}^{(l)} = \ln[1 + \exp(\boldsymbol{h}_{i,j}^{(l)'} \boldsymbol{W}_3^{(l)} + \boldsymbol{b}_3^{(l)})],
\end{aligned}
\tag{4}
$$

where $\boldsymbol{h}_{i,j}^{(0)} = \boldsymbol{x}_{i,j}$, $\{\boldsymbol{h}_{i,j}^{(l)}\}_{i=1,j=1,l=1}^{M,N,L} \in \mathbb{R}^{K_l}$, $\text{Re}LU(\cdot) = \max(0, \cdot)$ is the nonlinear activation function, and $\oplus$ denotes the concatenation in feature dimension.

Task latent variable inference For the same reason, we also use Weibull distributions to approximate the posteriors of task-specific latent variables $\{\boldsymbol{c}_i^{(l)}\}_{l=1}^L$, formulated as

$$
q(\boldsymbol{c}_i^{(l)} \,|\, \boldsymbol{x}_i, \boldsymbol{c}_j^{(l+1)}) = \text{Weibull}\left(\boldsymbol{k}_i^{(l)}, \boldsymbol{\lambda}_i^{(l)}\right), \tag{5}
$$

where the parameters $\boldsymbol{k}_i^{(l)}, \boldsymbol{\lambda}_i^{(l)} \in \mathbb{R}_+^{K_l}$ are deterministically transformed from both the observed task representation and

the information from the stochastic up-down path $\boldsymbol{c}_i^{(l+1)}$, as displayed in Fig. 2(b). In detail, we first employ a simple and effective permutation invariant neural network model (Zaheer et al., 2017) shown in Fig. 2(c) to infer hierarchical task representations as

$$\boldsymbol{h}_i^{(l)} = \frac{1}{N}\sum_{j=1}^{N} \boldsymbol{h}_{i,j}^{(l)}, \tag{6}$$

And then we calculate the parameters $\boldsymbol{k}_i^{(l)}, \boldsymbol{\lambda}_i^{(l)}$ by

$$\begin{aligned} \boldsymbol{h}_i^{(l)'} &= \begin{cases} \boldsymbol{h}_i^{(L)} & l = L, \\ \boldsymbol{h}_i^{(l)} \oplus \boldsymbol{\psi}^{(l+1)} \boldsymbol{c}_i^{(l+1)} & l < L, \end{cases} \\ \boldsymbol{k}_i^{(l)} &= \ln[1 + \exp(\boldsymbol{h}_i^{(l)'} \boldsymbol{W}_4^{(l)} + \boldsymbol{b}_4^{(l)})], \\ \boldsymbol{\lambda}_i^{(l)} &= \ln[1 + \exp(\boldsymbol{h}_i^{(l)'} \boldsymbol{W}_5^{(l)} + \boldsymbol{b}_5^{(l)})]. \end{aligned} \tag{7}$$

Additionally, Gaussian distributions are used to model potential uncertainties of the topics. Therefore, the inference of task-specific topic embeddings should consider both the likelihood information from observed task representations $\boldsymbol{h}_i^{(l)}$ and the prior information from $\boldsymbol{\beta}^{(l)}$, and we have

$$\begin{aligned} q(\boldsymbol{\alpha}_i^{(l)} \mid \boldsymbol{X}) &= \mathcal{N}\left(\boldsymbol{\mu}_{\boldsymbol{\alpha}_i}^{(l)}, \boldsymbol{\sigma}_{\alpha_i}^{(l)}\right), \\ \boldsymbol{\mu}_{\alpha_i}^{(l)} &= (\boldsymbol{h}_i^{(l)} \oplus \boldsymbol{\beta}^{(l)}) \boldsymbol{W}_6^{(l)} + \boldsymbol{b}_6^{(l)}, \\ \boldsymbol{\sigma}_{\alpha_i}^{(l)} &= (\boldsymbol{h}_i^{(t)} \oplus \boldsymbol{\beta}^{(l)}) \boldsymbol{W}_7^{(l)} + \boldsymbol{b}_7^{(l)}. \end{aligned} \tag{8}$$

Global latent variables inference The prior $\boldsymbol{\beta}^{(l)}$ captures the shared semantic structure of topic embedding across all the tasks, and we can use a Gaussian distribution to approximate its posterior as

$$q(\boldsymbol{\beta}^{(l)} \mid \boldsymbol{X}) = \mathcal{N}\left(\boldsymbol{\mu}_{\beta}^{(l)}, \boldsymbol{\sigma}_{\beta}^{(l)}\right), \tag{9}$$

where the inference network can be expressed as

$$\boldsymbol{\mu}_{\beta}^{(l)} = \boldsymbol{W}_8^{(l)}, \qquad \boldsymbol{\sigma}_{\beta}^{(l)} = \boldsymbol{W}_9^{(l)}. \tag{10}$$

Note that $\boldsymbol{\Omega} = \{\{\boldsymbol{W}_k^{(l)}\}_{k=1,t=1}^{9,L}, \{\boldsymbol{b}_k^{(l)}\}_{k=1,t=1}^{7,L}\}$ are the parameters of the inference network that can be seen as the shared structure representing the meta-knowledge (Gordon et al., 2018).

3.5. Variational inference under episodic training framework

As the definition described in Section 3.1, during evaluation, we are only given $\mathcal{D}_i^{(S)}$ to infer the variational distributions $q(\boldsymbol{\alpha}_i^{(l)})$ and $q(\boldsymbol{c}_i^{(l)})$ and measure the performance of the model by evaluating the variational distribution on corresponding $\mathcal{D}_i^{(Q)}$. In order to keep consistent during training and evaluation, we consider a modified version of the objective of the Evidence Lower Bound (ELBO) that incorporates this support and query task split (Ravi & Beatson, 2018). This means that for each episode i, we only use the support data $\boldsymbol{D}_i^{(S)}$ to infer the variational posteriors of $\left\{\boldsymbol{c}_i^{(l)}, \boldsymbol{\alpha}_i^{(l)}\right\}_{l=1}^{L}$, and the resulted objective is expressed as

$$\begin{aligned} L_{\text{ELBO}} = &\sum_{i=1}^{M} \mathbb{E}_Q \left[\ln p\left(\mathcal{D}_i | \left\{\boldsymbol{c}_i^{(l)}\right\}_{l=1}^{L}, \left\{\boldsymbol{\alpha}_i^{(l)}\right\}_{l=1}^{L}, \boldsymbol{\beta}^{(0)}\right)\right] \\ &- \sum_{i=1}^{M}\sum_{l=1}^{L} \mathbb{E}_Q \left[\ln \frac{q\left(\boldsymbol{c}_i^{(l)} | \mathcal{D}_i^{(S)}\right)}{p\left(\boldsymbol{c}_i^{(l)} | \boldsymbol{\psi}^{(l+1)} \boldsymbol{c}_i^{(l+1)}\right)}\right] \\ &- \sum_{i=1}^{M}\sum_{l=1}^{L} \lambda^{(t)} \mathbb{E}_Q \left[\ln \frac{q\left(\boldsymbol{\alpha}_i^{(t)} | \boldsymbol{D}_i^{(S)}\right)}{p\left(\boldsymbol{\alpha}_i^{(t)} | \boldsymbol{\beta}^{(l)}\right)}\right] - \sum_{l=1}^{L} \gamma^{(l)} \mathbb{E}_Q \left[\ln \frac{q\left(\boldsymbol{\beta}^{(l)}\right)}{p\left(\boldsymbol{\beta}^{(l)}\right)}\right], \end{aligned} \tag{11}$$

where $\lambda^{(l)}, \gamma^{(l)}$ are the regularization coefficients (Higgins et al., 2016), and the first item can be further written as

$$\begin{aligned} \mathbb{E}_Q \left[\ln p\left(\mathcal{D}_i | -\right)\right] = &\sum_{j=1}^{N} \mathbb{E}_Q \left[\ln p(\boldsymbol{x}_{i,j} | \boldsymbol{\Phi}_i^{(1)}, \boldsymbol{\theta}_{i,j}^{(1)})\right] \\ &- \sum_{j=1}^{N}\sum_{l=1}^{T} \mathbb{E}_Q \left[\ln \frac{q(\boldsymbol{\theta}_{i,j}^{(l)} \mid \boldsymbol{x}_{i,j}, \boldsymbol{\theta}_{i,j}^{(l+1)}, \boldsymbol{c}_j^{(l)})}{p(\boldsymbol{\theta}_{i,j}^{(l)} | \boldsymbol{\Phi}_i^{(l+1)}, \boldsymbol{\theta}_{i,j}^{(l+1)}, \boldsymbol{c}_i^{(l)})}\right], \end{aligned} \tag{12}$$

with

$$\begin{aligned} Q = \prod_{l=1}^{L} &q(\boldsymbol{\theta}_{i,j}^{(l)} \mid \boldsymbol{x}_{i,j}, \boldsymbol{\theta}_{i,j}^{(l+1)}, \boldsymbol{c}_j^{(l)}) \\ &q(\boldsymbol{c}_i^{(l)} | \mathcal{D}_i^{(S)}) q(\boldsymbol{\alpha}_i^{(l)} | \mathcal{D}_i^{(S)}) q(\boldsymbol{\beta}^{(l)}), \end{aligned} \tag{13}$$

where $q(\boldsymbol{\theta}_{i,j}^{(l)} \mid \boldsymbol{x}_{i,j}, \boldsymbol{\theta}_{i,j}^{(l+1)}, \boldsymbol{c}_j^{(l)})$ is the variational Weibull distribution in Eq. (3), and $p(\boldsymbol{\theta}_{i,j}^{(l)})$ is the prior gamma distribution in Eq. (1); The first term is the expected log-likelihood or reconstruction error, while the second term is the Kullback–Leibler (KL) divergence that constrains $q(\boldsymbol{\theta}_j^{(l)} \mid \boldsymbol{x}_{i,j}, \boldsymbol{\theta}_{i,j}^{(l+1)}, \boldsymbol{c}_j^{(l)})$ to be close to its prior $p(\boldsymbol{\theta}_j^{(l)})$ in

Algorithm 1 Autoencoding Variational Inference for Meta-DETM

Set mini-batch size m and the number of layer T
Initialize the variational network parameters $\boldsymbol{\Omega}$;
while not converge **do**
1. Randomly sample a mini-batch of m support and query sets to form a subtask $\left\{\mathcal{D}_i^{(S)}\right\}_{i=1}^{m} \left\{\mathcal{D}_i^{(Q)}\right\}_{i=1}^{m}$;
2. Infer variational posterior for task latent variables $\{\boldsymbol{\alpha}_i^{(l)}\}_{i=1,l=1}^{m,L}$ and $\{\boldsymbol{c}_i^{(l)}\}_{i=1,l=1}^{m,L}$ only using $\left\{\mathcal{D}_i^{(S)}\right\}_{1,m}$ by Eq. (5) and Eq. (8);
3. Infer variational posterior for document latent variables $\{\boldsymbol{\theta}_{i,j}^{(l)}\}_{i=1,l=1,j=1}^{m,L,N}$ using $\{\mathcal{D}_{i,j}\}_{i=1,j=1}^{m,N}$ by Eq. (3);
4. Calculate $\nabla_{\boldsymbol{\Omega}} L\left(\boldsymbol{\Omega}; \left\{\mathcal{D}_i^{(S)}\right\}_{i=1}^{m}\right)$ according to Eq. (11) and update $\boldsymbol{\Omega}$.
end while

the generative model; The analytic KL expression and simple reparameterization of the Weibull distribution make it simple to estimate the gradient of the ELBO, with respect to all the parameters $\mathbf{\Omega}$ in the inference network. Similar to variational auto-encoders (VAEs) (Kingma & Welling, 2013; Rezende et al., 2014), the training objective of Meta-DETM is to maximize the ELBO.

3.6. Model properties

Hierarchical semantic topic: The loading matrices $\mathbf{\Phi}^{(l)}$ in Eq. (2) capture the semantic correlations of the topics of adjacent layers. Using the law of total expectation, we have

$$\mathbb{E}\left[\boldsymbol{x}_j|\boldsymbol{\theta}_j^{(l)}, \left\{\mathbf{\Phi}^{(t)}, e_j^{(t)}\right\}_{t=1}^{l}\right] = \left[\prod_{t=1}^{l}\mathbf{\Phi}^{(t)}\right]\frac{\boldsymbol{\theta}_j^{(l)}}{\prod_{t=2}^{l} e_j^{(t)}}. \quad (14)$$

Therefore, $\left[\prod_{t=1}^{l-1}\mathbf{\Phi}^{(t)}\right]\mathbf{\Phi}^{l}$ is naturally interpreted as the projection of topic $\mathbf{\Phi}^{(l)}$ to the observation space, providing us with a principled way to visualize the topics at different semantic levels.

Multi-level semantic task dependence: The proposed model can capture semantic task dependences in different levels by hierarchical task-specific latent variables $\{\boldsymbol{c}^{(l)}\}_{l=1}^{L}$, which represent more specific semantic structure dependence in the lower layer and more abstract semantic structure dependence in the higher layer.

Effectiveness of meta-decoder: Our designed meta-decoder enables knowledge accumulated from source tasks to be efficiently transferred to the target task, yielding a good posterior approximation with only a few samples.

4. Experiments

We carry out extensive experiments on several widely-used text datasets varying in scale. We report per-heldout-word perplexity and few-shot text classification accuracy on each dataset. We also qualitatively analyze the adaptation process from shared global topics to task-specific topics. **Datasets:** Our experiments are conducted on three widely-used textual benchmarks with different scales and document lengths, including 20Newsgroups (20NG), Yahoo! Answers (Yahoo), and Reuters Corpus Volume I (RCV1). 20NG consists of informal discourse from news discussion forums and has 18,846 documents from 20 categories (Lang, 1995). Yahoo is a topic classification dataset built from Yahoo! Answers Comprehensive Questions and Answers version 1.0 dataset, with 10 categories, each containing 145,000 samples (Zhang et al., 2015). We randomly sample 100,000 documents from the Yahoo dataset for our experiments. RCV1, a collection of Reuters news wire articles from 1996 to 1997, consists of 804,414 documents (Lewis et al., 2004). These articles are written in formal speech and labeled with topic codes. We consider 55 second-level topics as our total class set for our experiments. We build a vocabulary by removing the stop words and selecting the top K most frequent words, where K = 2,000 for 20NG, K= 10,000 for RCV1, and K= 10,000 for Yahoo in the per-heldout-word perplexity (PPL) experiments, and K = 2,000 for both 20NG and RCV1 for few-shot document classification experiments.

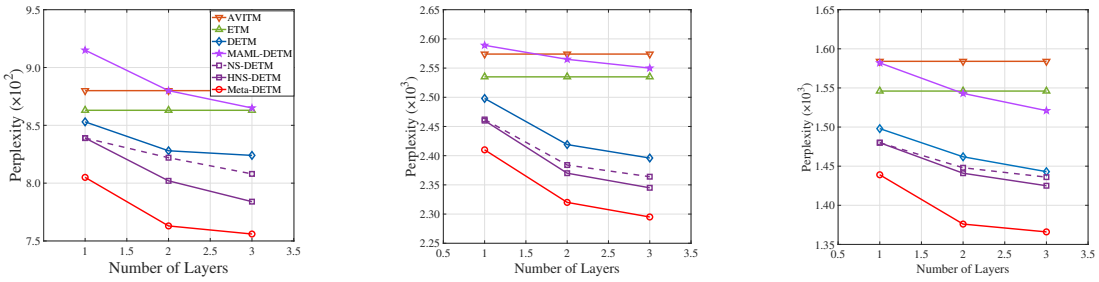
Baseline methods: We compare the proposed model with classical neural topic models and their few-shot variants, including: Latent Dirichlet allocation (LDA) with Products of Experts (AVITM) (Srivastava & Sutton, 2017), which replaces the mixture model in LDA with a product of experts and uses the AVI for training; Embedded topic Model (ETM) (Dieng et al., 2020), a variant of LDA that incorporates the idea of word embeddings; Sawtooth Factorial Embedded Topic Model (DETM) (Duan et al., 2021a), which extends the gamma belief network (Zhou et al., 2015) to a deep embedded topic model by taking inspiration from ETM. For all the above baselines, the model is trained on all training corpora and then fine-tuned with the support set from the target task at the test time. Besides, we also consider three few-shot variants for ablation study, including MAML-DETM, which optimizes all the parameters of DETM using model-agnostic meta-learning (Finn et al., 2017); neural statistician (Edwards & Storkey, 2016) for DETM (NS-DETM), which only employs task-specific latent variables at the highest layer to modulate generative model; and its extension hierarchical neural statistician for DETM (Giannone & Winther, 2021) (HNS-DETM), which uses hierarchical task-specific latent variables to modulate the generative model.

4.1. Per-heldout-word perplexity

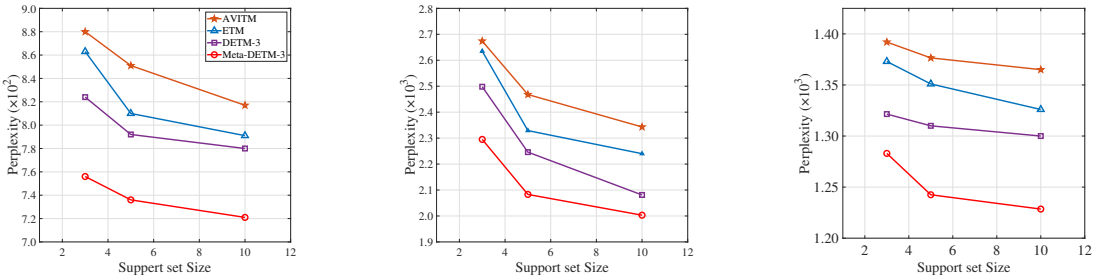
Training/test split: For 20NG, we select 18 classes as the training corpora and the remaining two classes as the validation corpus and the test corpus, respectively; For Yahoo, 8 classes are chosen to form the training corpora, another two classes are used as the validation corpus and the test corpus, respectively. For RCV1, we split the total 55 classes into 47, 3, 5 to comprise the training, validation, and test corpus.

Model setting: For hierarchical topic models, we set the network structure with three layers as [256, 128, 64]. For embedded topic models such as ETM, DETM, NS-DETM, HNS-DETM, and Meta-DETM, we set the embedding size as 50. For the NTMs, we set the hidden size as 256. For optimization, the Adam optimizer (Kingma & Ba, 2014) is adopted with an initial learning rate of $1e^{-2}$. We set support size as [3,5,10], and query size as 15, mini-batch size is defined as the class number of the training corpus. For a testing time, we sample tasks from testing corpus with support size [3,5,10] and query size [3,5,10], and average PPL for all the query documents to get final test results.

Results: To evaluate the predictive performance of the proposed model, we calculate the per-heldout-word perplexity



(a) 20NG (b) Yahoo (c) RCV1



(d) 20NG (e) Yahoo (f) RCV1

Figure 3. (a)-(c): Comparison of per-heldout-word perplexity with set size 3. (d)-(f): Comparison of per-heldout-word perplexity with different support set size [3, 5, 10] (Lower test perplexity indicates better performance of the estimated topic model).

(PPL) (Cong et al., 2017) on three regular document datasets, including 20NG, Yahoo, and RCV1. During testing, we first use the support set to estimate a topic model for the target task. And then for each document in the query set, we randomly select 80% of the word tokens to form a training matrix X, holding out the remaining 20% to form a testing matrix Y. We use X to infer the latent variables and calculate the per-held-word perplexity as

$$\exp\left\{-\frac{1}{y_{\cdot\cdot}}\sum_{v=1}^{V}\sum_{n=1}^{N} y_{vn}\ln\frac{\sum_{s=1}^{S}\sum_{k=1}^{K^1}\phi_{vk}^{(1)s}\theta_{kn}^{(1)s}}{\sum_{s=1}^{S}\sum_{v=1}^{V}\sum_{k=1}^{K^1}\phi_{vk}^{(1)s}\theta_{kn}^{(1)s}}\right\},$$

where S is the total number of collected samples and $y_{\cdot\cdot} = \sum_{v=1}^{V}\sum_{n=1}^{N} y_{vn}$. Fig. 3 (a)-(c) show how the perplexity changes as a function of the number of layers for various models over three different datasets. Benefiting from the powerful embedding-based decoder, ETM performs better than AVITM. But both are constrained by the shallow structure that lacks expressiveness, resulting in a performance gap with the deep embedding topic models. With a similar embedding-based decoder, we can see that DETM with a single hidden layer outperforms ETM, indicating that using the Weibull distribution is more appropriate than the logistic normal distribution to model the latent document representation. Although equipped with a powerful meta-learning algorithm to learn good initialization for few-shot adaption, MAML-DETM yields worse perplexity than its base DETM. This may be due to the generalization difficulty of using only a few samples to compute gradients in a high-dimensional parameter space (Rusu et al., 2018). By learning a good task-specific prior for the document-specific factor score, NS-DETM gets certain performance improvements on DETM. Furthermore, HNS-DETM gets a more noticeable performance boost by modeling the hierarchical task-specific variables as prior information for the document-specific factor score across all layers. Finally, equipped with an embedding-based topic meta-learner, Meta-DETM outperforms the other methods by a large margin. Fig. 3 (d)-(f) show how the perplexity changes as a function of support set size for various models over three different datasets. It's not surprised that the performance becomes better as the set size increases, and the proposed Meta-DETM gets the best performance compared with other base models.

Model	Set size	20News	Yahoo	Rcv1
LDA	3	1751	4984	9932
LDA	5	1670	4710	9147
LDA	10	1523	3808	8789
GBN-1	1	1692	4943	9954
GBN-1	5	1633	4681	8756
GBN-1	10	1489	3765	8553
DETM-1	1	853	1498	2497
DETM-1	5	813	1382	2414
DETM-1	10	768	1345	2350

Table 1. PPL results on three datasets.

We also evaluate the performance of traditional topic models such as LDA and GBN, the results are presented in Tab. 1. From the table we can see, traditional topic models like

LDA and GBN perform worse than the neural topic model. And the relevant results can be found in Iwata (2021), compared with trained individually using the target support data, fine-tuning pre-trained topic models on the training corpus by target support data gets worse performance. The results show that the traditional topic models are hard to be adapted to the new task with a few samples, and using an unsuitable pre-trained model as a prior has a bad influence on meta-learning. On the other hand, the experiment results further demonstrate the continuous learning ability of traditional topic models in that the topic is hard to change (corresponding to catastrophic forgetting) (Chen et al., 2021).

4.2. Few-shot text classification

To further validate the effectiveness of the proposed model, we also compare our model with several baseline algorithms on the few-shot text classification task.

Training/test split: Since each of the 20 categories in 20NG can be further grouped into 6 higher-level categories (*i.e.*, *computers*, *recreation*, *science*, *politics*, *religion*, and *forsale*), we conduct the split to ensure that no higher-level category spans across different split sets. For instance, we choose those categories belonging to "*science*" and "*recreation*" to form the training corpus, categories attached to "*computers*" as the validation corpus, and all other categories as the test corpus. For the RCV1 dataset, we first sample 20 documents from each class to obtain a small version of RCV1 dataset. Then we divide the 55 classes into a training set of 35 classes, a validation set of 5 classes, and a test set of 15 classes.

Model setting: We evaluate the model performance under two types of representations with three typical few-shot learning algorithms. MLP applies a two-layer feed-forward network over the input bag-of-words features. CNN applies 1D convolution over the input words and obtains the representation by max-over-time pooling (Kim, 2014). Here we don't use the transformer architecture to get document representations in the light of its huge amount of parameters, which is not friendly for few-shot learning. FT first pre-trains a classifier over all training examples, then fine-tunes the network using the support set (Chen et al., 2019). MAML learns an initialization for all the model parameters, so that the model can quickly adapt to new classes (Finn et al., 2017). Prototypical network (PROTO) learns a metric space for few-shot classification by minimizing the Euclidean distance between the centroid of each class and its constituent examples. The 20NG and RCV1 datasets are used with a vocabulary of size 2,000. The network structure of the neural deep topic model is set as [128, 64, 32]. The support set size is set as [1, 5], the query size is set as 15, and the mini-batch size is set as the class number of the training corpus. Other model settings are kept consistent with the PPL experiments..

Method		20 News		RCV1	
Rep.	Alg.	1 shot	5 shot	1 shot	5 shot
MLP	FT	31.5	40.2	56.1	68.2
CNN	FT	29.0	36.1	55.0	62.3
MLP	MAML	29.2	36.2	50.2	60.3
CNN	MAML	28.4	35.7	46.0	56.1
MLP	PROTO	29.5	36.1	39.8	50.7
CNN	PROTO	28.6	33.0	37.4	48.3
HNS-DETM-1		30.3	39.2	51.8	63.0
HNS-DETM-3		32.5	42.5	53.2	66.6
Meta-DETM-1		33.7	44.1	54.4	66.4
Meta-DETM-3		**34.4**	**45.5**	**57.2**	**72.3**

Table 2. Results of 5-way 1-shot and 5-way 5-shot classification on two datasets.

Results: We first describe in detail how to finish the classification task with our model. Specifically, given a set of labelled examples of each unseen class $\mathcal{D}_0, ..., \mathcal{D}_4$, we first compute the approximate posteriors $q(\{\boldsymbol{c}\}_{t=1}^T|\mathcal{D}_i;\boldsymbol{\Omega})$ and $q(\{\alpha\}_{t=1}^T|\mathcal{D}_i;\boldsymbol{\Omega})$. Then for each sample x, we compute the test ELBO by Eq. 12 with the approximate posteriors of different class sets and classify it according to maximizing the test ELBO. Table 2 lists the comparison of various few-shot learning algorithms on two real-world datasets, including 20 News and Rcv1. Meta-learning approaches such as MAML and PROTO have emerged as promising methods for few-shot image classification. However, different from vision models that can share low-level patterns (such as edges) and their corresponding representations across tasks, highly informative words for one task may not be relevant for other tasks (such as the word 'grandma' should be informative in the family class but not in the internet class) (Bao et al., 2019). So it's not surprising that the meta-learner's performance drops below that of a simple, fine-tuned method. From another principle, the proposed models meta-learn how to build a topic model for each task and hence need not face the above challenge. In detail, we see that both HNS-DETM and Meta-DETM get strong performance compared with other models. Further, Meta-DETM gets the best performance, illustrating the effectiveness of the embedded topic meta-learner. Besides, the performance of Meta-DETM improves as the layer gets deeper, demonstrating the significance of exploring hieratical semantic structure. Overall, the Meta-DETM can be a strong classifier and provides a new consideration for few-shot text classification.

4.3. Effect of the regularization coefficient

To generate topic embeddings for a new task, the embedded topic meta-learner mainly fuses the information from i) the global topic embeddings learned from a large number

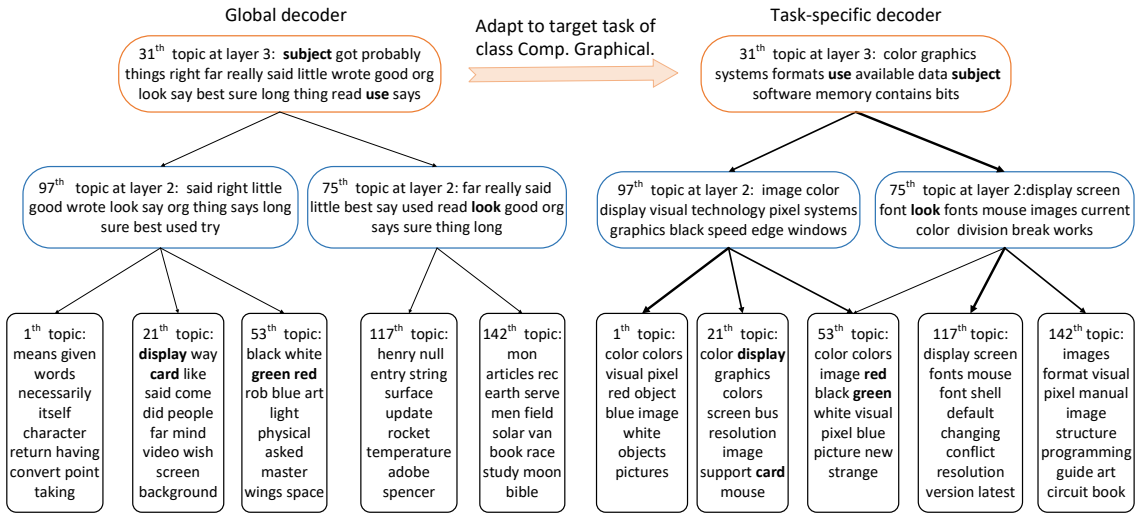


Figure 4. Hierarchical topics learned for the target task of class Comp.Graphical (right) and its corresponding global topic (left).

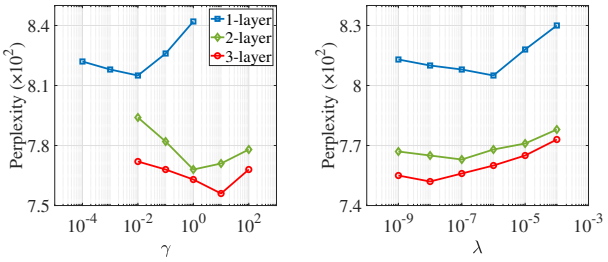


Figure 5. The effect of (a) $\gamma^{(t)}$ and (b) $\lambda^{(t)}$ on model's performance.

of training tasks and ii) the task representations inferred from the support set. A primary concern in meta-learning is how to balance the two pieces of information to achieve better generalization (Yin et al., 2019). In our model, as defined in Eq. (11), $\mathbb{E}[\ln q(\boldsymbol{\beta}^{(t)})/p(\boldsymbol{\beta}^{(t)})]$ can be regarded as a meta-regularization to the meta-parameters $\boldsymbol{\beta}^{(t)}$, which limits the information from the training tasks stored in the meta-parameters; and $\mathbb{E}[\ln q(\alpha_i^{(t)}|\mathcal{D}_i^{(S)})/p(\boldsymbol{\alpha}_i^{(t)}|\boldsymbol{\beta}^{(t)})]$ can be seen as the regularization to the parameters $\boldsymbol{\alpha}^{(t)}$, which limits the information from the new task. To study the effect of the regularization coefficient $\gamma^{(t)}$ and $\lambda^{(t)}$ in Eq. (11), we evaluate the test perplexity with a varying layer-wise regularization coefficient on the 20NG dataset. The results of related experiments are shown in Fig. 5, it indicates that a better balance between the two pieces of information can improve performance. Furthermore, we can see that the optimal value of γ becomes larger as the layer goes deeper, which shows higher layers require less historical information. Meanwhile, the optimal value of λ is smaller in the deeper layers, indicating that higher layers need more new task information. Such phenomena well support our motivation that low-level (specific) structures learned by shallow layers can be shared across various tasks, whereas deeper layers cannot.

4.4. Hierarchical semantic topics adaption

In this part, we visually show the learned topics at different layers. As shown in Fig. 4, we exhibit hierarchical topics adapted for a new task from the Comp.Graphical class (right) and the corresponding global topics learned from historical training tasks (left) on the 20NG dataset. For the 1th layer, some task-specific topics have more similar semantics with global topics, such as the 21th and 53th topics, and we can see that the task-specific topics are only slightly adjusted compared to the corresponding global topics. Meanwhile, some other global topics that are not related with the new incoming task will undergo more significant changes, such as the 1th and 142th topics. For the deeper layers, most task-specific topics have different semantics with the global topics, showing that at higher layers topics depend more on information from the new task. This phenomenon further confirms the investigation in Sec. 4.3.

5. Conclusion

In this paper, we have studied the problem of few-shot learning for topic modeling, which has not yet attracted much research attention. We propose a novel framework that can efficiently extend embedded topic models to discover new topics from only a few documents. Extensive experiments demonstrate that our proposed framework is more effective than several typical few-shot learning algorithms. Moreover, we also show how our model finds new topics from a few documents by visualization, which helps us to better understand the mechanism of our model's effectiveness.

Acknowledgments

Bo Chen acknowledges the support of NSFC (U21B2006 and 61771361), Shaanxi Youth Innovation Team Project, the 111 Project (No. B18039) and the Program for Oversea Talent by Chinese Central Government.

References

Bao, Y., Wu, M., Chang, S., and Barzilay, R. Few-shot text classification with distributional signatures. *arXiv preprint arXiv:1908.06039*, 2019.

Bianchi, F., Terragni, S., and Hovy, D. Pre-training is a hot topic: Contextualized document embeddings improve topic coherence. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pp. 759–766, aug 2021.

Blei, D. M., Ng, A. Y., and Jordan, M. I. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan): 993–1022, 2003.

Blei, D. M., Griffiths, T. L., and Jordan, M. I. The nested chinese restaurant process and bayesian nonparametric inference of topic hierarchies. *Journal of the ACM (JACM)*, 57(2):1–30, 2010.

Card, D., Tan, C., and Smith, N. A. A neural framework for generalized topic models. *arXiv preprint arXiv:1705.09296*, 2017.

Chen, W., Chen, B., Liu, Y., Cao, X., Zhao, Q., Zhang, H., and Tian, L. Max-margin deep diverse latent dirichlet allocation with continual learning. *IEEE Transactions on Cybernetics*, 2021.

Chen, W.-Y., Liu, Y.-C., Kira, Z., Wang, Y.-C. F., and Huang, J.-B. A closer look at few-shot classification. *arXiv preprint arXiv:1904.04232*, 2019.

Cong, Y., Chen, B., Liu, H., and Zhou, M. Deep latent dirichlet allocation with topic-layer-adaptive stochastic gradient riemannian mcmc. *arXiv preprint arXiv:1706.01724*, 2017.

Dieng, A. B., Ruiz, F. J., and Blei, D. M. Topic modeling in embedding spaces. *Transactions of the Association for Computational Linguistics*, 8:439–453, 2020.

Duan, Z., Wang, D., Chen, B., Wang, C., Chen, W., Li, Y., Ren, J., and Zhou, M. Sawtooth factorial topic embeddings guided gamma belief network. In *International Conference on Machine Learning*, pp. 2903–2913. PMLR, 2021a.

Duan, Z., Xu, Y., Chen, B., Wang, D., Wang, C., and Zhou, M. Topicnet: Semantic graph-guided topic discovery. In *Thirty-Fifth Conference on Neural Information Processing Systems*, 2021b.

Edwards, H. and Storkey, A. Towards a neural statistician. *arXiv preprint arXiv:1606.02185*, 2016.

Finn, C., Abbeel, P., and Levine, S. Model-agnostic meta-learning for fast adaptation of deep networks. In *International Conference on Machine Learning*, pp. 1126–1135. PMLR, 2017.

Gan, Z., Chen, C., Henao, R., Carlson, D., and Carin, L. Scalable deep poisson factor analysis for topic modeling. In *International Conference on Machine Learning*, pp. 1823–1832, 2015.

Giannone, G. and Winther, O. Hierarchical few-shot generative models. *arXiv preprint arXiv:2110.12279*, 2021.

Gordon, J., Bronskill, J., Bauer, M., Nowozin, S., and Turner, R. E. Meta-learning probabilistic inference for prediction. *arXiv preprint arXiv:1805.09921*, 2018.

Guo, D., Chen, B., Zhang, H., and Zhou, M. Deep Poisson gamma dynamical systems. *arXiv preprint arXiv:1810.11209*, 2018.

Gupta, P., Chaudhary, Y., Runkler, T., and Schütze, H. Neural topic modeling with continual lifelong learning. In *International Conference on Machine Learning*, pp. 3907–3917. PMLR, 2020.

Henao, R., Gan, Z., Lu, J., and Carin, L. Deep poisson factor modeling. *Advances in Neural Information Processing Systems*, 28:2800–2808, 2015.

Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., and Lerchner, A. beta-vae: Learning basic visual concepts with a constrained variational framework. 2016.

Hoffman, M. D., Blei, D. M., Wang, C., and Paisley, J. Stochastic variational inference. *Journal of Machine Learning Research*, 14(5), 2013.

Hu, C., Rai, P., and Carin, L. Transfer learning for hierarchically supervised topic models. In *2015 NIPS Workshop in Transfer and Multi-task Learning, Montreal, Canada*, 2015.

Huynh, V., Zhao, H., and Phung, D. Otlda: A geometry-aware optimal transport approach for topic modeling. *Advances in Neural Information Processing Systems*, 33: 18573–18582, 2020.

Iakovleva, E., Verbeek, J., and Alahari, K. Meta-learning with shared amortized variational inference. In *International Conference on Machine Learning*, pp. 4572–4582. PMLR, 2020.

Iwata, T. Few-shot learning for topic modeling. *arXiv preprint arXiv:2104.09011*, 2021.

Jeon, I., Park, Y., and Kim, G. Neural variational dropout processes. In *International Conference on Learning Representations*, 2022.

Kim, J., Kim, T., Kim, S., and Yoo, C. D. Edge-labeling graph neural network for few-shot learning. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11–20, 2019.

Kim, Y. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1746–1751, Doha, Qatar, October 2014. Association for Computational Linguistics. doi: 10.3115/v1/D14-1181. URL https://aclanthology.org/D14-1181.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Kingma, D. P. and Welling, M. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Lang, K. Newsweeder: Learning to filter netnews. In *Machine Learning Proceedings 1995*, pp. 331–339. Elsevier, 1995.

Lewis, D. D., Yang, Y., Rose, T. G., and Li, F. Rcv1: A new benchmark collection for text categorization research. *Journal of machine learning research*, 5(Apr):361–397, 2004.

Miao, Y., Yu, L., and Blunsom, P. Neural variational inference for text processing. In *International conference on machine learning*, pp. 1727–1736, 2016.

Mimno, D., Wallach, H., Naradowsky, J., Smith, D. A., and McCallum, A. Polylingual topic models. In *Proceedings of the 2009 conference on empirical methods in natural language processing*, pp. 880–889, 2009.

Mishra, N., Rohaninejad, M., Chen, X., and Abbeel, P. A simple neural attentive meta-learner. *arXiv preprint arXiv:1707.03141*, 2017.

Oord, A., Vinyals, O., and Kavukcuoglu, K. Neural discrete representation learning. 2017.

Paisley, J., Wang, C., Blei, D. M., and Jordan, M. I. Nested hierarchical dirichlet processes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(2):256–270, 2014.

Ravi, S. and Beatson, A. Amortized bayesian meta-learning. In *International Conference on Learning Representations*, 2018.

Ravi, S. and Larochelle, H. Optimization as a model for few-shot learning. 2016.

Rezende, D. J., Mohamed, S., and Wierstra, D. Stochastic backpropagation and variational inference in deep latent gaussian models. In *International Conference on Machine Learning*, volume 2, 2014.

Rubin, T. N., Chambers, A., Smyth, P., and Steyvers, M. Statistical topic models for multi-label document classification. *Machine learning*, 88(1-2):157–208, 2012.

Rusu, A. A., Rao, D., Sygnowski, J., Vinyals, O., Pascanu, R., Osindero, S., and Hadsell, R. Meta-learning with latent embedding optimization. *arXiv preprint arXiv:1807.05960*, 2018.

Santoro, A., Bartunov, S., Botvinick, M., Wierstra, D., and Lillicrap, T. Meta-learning with memory-augmented neural networks. In *International conference on machine learning*, pp. 1842–1850. PMLR, 2016.

Snell, J., Swersky, K., and Zemel, R. S. Prototypical networks for few-shot learning. *arXiv preprint arXiv:1703.05175*, 2017.

Srivastava, A. and Sutton, C. Autoencoding variational inference for topic models. *arXiv preprint arXiv:1703.01488*, 2017.

Sung, F., Yang, Y., Zhang, L., Xiang, T., Torr, P. H., and Hospedales, T. M. Learning to compare: Relation network for few-shot learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1199–1208, 2018.

Vinyals, O., Blundell, C., Lillicrap, T., Wierstra, D., et al. Matching networks for one shot learning. *Advances in neural information processing systems*, 29:3630–3638, 2016.

Wang, D., Guo, D., Zhao, H., Zheng, H., Tanwisuth, K., Chen, B., and Zhou, M. Representing mixtures of word embeddings with mixtures of topic embeddings. In *ICLR 2022: International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=IYMuTbGzjFU.

Wang, X., McCallum, A., and Wei, X. Topical n-grams: Phrase and topic discovery, with an application to information retrieval. In *Seventh IEEE International Conference on Data Mining (ICDM 2007)*, pp. 697–702, 2007.

Yin, M., Tucker, G., Zhou, M., Levine, S., and Finn, C. Meta-learning without memorization. *arXiv preprint arXiv:1912.03820*, 2019.

Zaheer, M., Kottur, S., Ravanbakhsh, S., Poczos, B., Salakhutdinov, R., and Smola, A. Deep sets. *arXiv preprint arXiv:1703.06114*, 2017.

Zhang, H., Chen, B., Guo, D., and Zhou, M. Whai: Weibull hybrid autoencoding inference for deep topic modeling. *arXiv preprint arXiv:1803.01328*, 2018.

Zhang, H., Chen, B., Cong, Y., Guo, D., Liu, H., and Zhou, M. Deep autoencoding topic model with scalable hybrid bayesian inference. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

Zhang, X., Zhao, J., and LeCun, Y. Character-level convolutional networks for text classification. *Advances in neural information processing systems*, 28:649–657, 2015.

Zhao, H., Du, L., Buntine, W., and Zhou, M. Dirichlet belief networks for topic structure learning. *arXiv preprint arXiv:1811.00717*, 2018.

Zhao, H., Phung, D., Huynh, V., Le, T., and Buntine, W. Neural topic model via optimal transport. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=Oos98K9Lv-k.

Zhou, M., Cong, Y., and Chen, B. The poisson gamma belief network. *Advances in Neural Information Processing Systems*, 28:3043–3051, 2015.

Zhou, M., Cong, Y., and Chen, B. Augmentable gamma belief networks. *The Journal of Machine Learning Research*, 17(1):5656–5699, 2016.