Bayesian Progressive Deep Topic Model with Knowledge Informed Textual Data Coarsening Process

Zhibin Duan [*1] **Xinyang Liu** [*1] **Yudi Su** [1] **Yishi Xu** [1] **Bo Chen** [1] **Mingyuan Zhou** [2]

Abstract

Deep topic models have shown an impressive ability to extract multi-layer document latent representations and discover hierarchical semantically meaningful topics. However, most deep topic models are limited to the single-step generative process, despite the fact that the progressive generative process has achieved impressive performance in modeling image data. To this end, in this paper, we propose a novel progressive deep topic model that consists of a knowledge-informed textural data coarsening process and a corresponding progressive generative model. The former is used to build multi-level observations ranging from concrete to abstract, while the latter is used to generate more concrete observations gradually. Additionally, we incorporate a graph-enhanced decoder to capture the semantic relationships among words at different levels of observation. Furthermore, we perform a theoretical analysis of the proposed model based on the principle of information theory and show how it can alleviate the well-known "latent variable collapse" problem. Finally, extensive experiments demonstrate that our proposed model effectively improves the ability of deep topic models, resulting in higher-quality latent document representations and topics.

1. Introduction

Topic modeling has developed into one of the most widely-used techniques for text analysis. Bayesian probabilistic topic models (PTMs), such as latent Dirichlet allocation (LDA) (Blei et al., 2003) and Poisson factor analysis (PFA) (Zhou et al., 2012), are built on the assumption that each document is represented by a mixture of topics, where each topic defines a probability distribution over words and describes an interpretable semantic concept. Besides, these models can also derive low-dimensional representations of the documents, which have proven useful in a series of natural language processing tasks (Rubin et al., 2012; Wang et al., 2007; Mimno et al., 2009).

While these shallow topic models are widely used, their modeling ability is still limited by the single-layer structure, which makes it difficult to explore the hierarchical semantic structure (Marius & Burkhardt). To this end, there has been an emerging research interest in building deep topic models (DTMs) (Blei et al., 2010; Paisley et al., 2014; Gan et al., 2015; Zhou et al., 2016; Zhao et al., 2018) that aims to mine multi-layer document representations and discover meaningful topic taxonomies. Recently, the success of deep generative models such as variational autoencoder (VAE) (Kingma & Welling, 2013; Rezende et al., 2014) has shown the potential of deep neural networks in posterior inference, motivating the proposal of a range of neural topic models (NTMs) ranging from shallow structure (Srivastava et al., 2017; Miao et al., 2017) to deep structure (Zhang et al., 2018). Compared with Bayesian PTMs, NTMs usually enjoy better flexibility and scalability, which are essential for applications on large-scale data and downstream tasks.

Despite considerable effort has been put into developing more effective DTMs, most of them rely on a single-step generative process. The counterpart to this is the progressive generative models that synthesize images in a coarse-to-fine manner, which have attracted wide attention due to their impressive performance (Karras et al., 2017; Razavi et al., 2019; Ho et al., 2020; Austin et al., 2021; Bansal et al., 2022; Shu & Ermon, 2022; Lee et al., 2022; Gu et al., 2022). Meanwhile, some works (Shen et al., 2019; Tan et al., 2020) have attempted to build progressive language models for long sequence generation, which show that progressive generative models have the potential to model more complex data distribution (longer sequences). Further, it should be emphasized that the majority of language models use an autoregressive technique for generating sequences (Radford et al., 2019), which can be regarded as a progressive genera-

[*]Equal contribution [1]National Key Laboratory of Radar Signal Processing, Xidian University, Xi'an, 710071, 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 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

tive process that generates longer sequences gradually; and the non-autoregressive language models, which generate all of the words in a document in a single step, usually perform worse than the autoregressive language models (Xiao et al., 2023). Like the generative process of the non-autoregressive language models, the original topic model generates Bag-of-words in a single step. Overall, topic modeling has an appealing potential to enhance its modeling capability by applying a progressive generation approach.

For how to build a progressive generative process, we get inspired from the analogy of topic hierarchy and knowledge graph to propose a knowledge informed textual data coarsening process. Specifically, the progressive generative models mainly consist of a forward process that coarsens an image gradually by downsampling or blurring, followed by a corresponding reverse process (generative process) that upsamples or deblurs progressively. To the best of our knowledge, most progressive generative models mainly focus on image data. One reason is that the image coarsening process can be naturally achieved by pooling the neighboring pixels in this space due to images having semantic consistency in Euclidean space. However, this method cannot be directly transferred to the process of text coarsening, as the semantic dependency between words in text is more complex than the spatial dependency. This leaves a challenge in building a progressive generative process for text data.

Fortunately, we find that external knowledge, such as knowledge graph, can be used to measure the semantic relationships between two words. As shown in Fig. 1(a), words can be organized as a concept hierarchy with hypernym relations (Miller, 1995), and the parent nodes have more abstract semantics and contain the semantics of their child nodes. In the concept hierarchy, the nodes at different layers can be regarded as having different levels of semantics, and the nodes at higher layers have more abstract semantics. For example, the concept "organization" has more abstract semantics compared with "company" and "university". Inspired with the semantic structure above, we develop a general framework for gradually coarsening textual data from concrete to abstract, which will be described in Sec.3.1 in detail. And this framework will serve as the forward process in a progressive generative process for textual data.

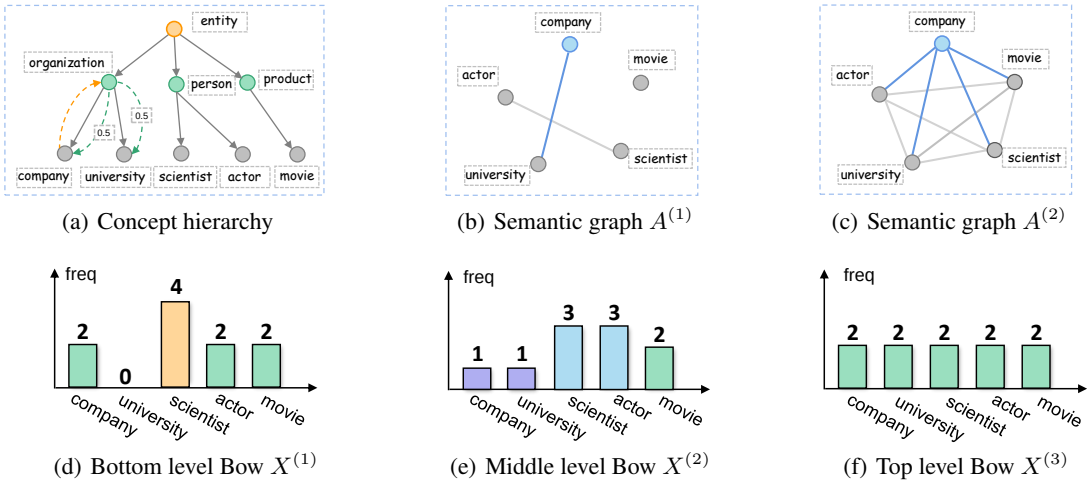
Motivated by the former works on the progressive model and deep topic models, we formulate a novel progressive generative model, named progressive gamma belief network (ProGBN), which models text data in a coarse-to-fine manner. The proposed ProGBN can be seen as the corresponding reverse process in a progressive generative process, progressively generating more concrete texture data. Meanwhile, considering the textual data coarsening process will establish new semantic dependencies among words, we develop a graph-enhanced decoder, which can capture this semantic dependence. After that, we designed a hierarchical inference network to approximate the posterior of the latent variables in ProGBN under the VAE framework (Kingma & Welling, 2013). Finally, to verify the benefits of the progressive generation process, we take theoretically analyze for the ProGBN from the perspective of information theory (Thomas & Joy, 2006). Our analysis reveal that ProGBN can effectively alleviate the well-known latent variable collapse issue (Dieng et al., 2019a; Li et al., 2022a) in hierarchical VAEs. We summarize our contributions as follows:

- A general knowledge-informed textual data coarsening process is developed that can coarsen text from concrete to abstract.
- A novel progressive deep topic model, equipped with a graph-enhanced decoder, are built to progressively generate more concrete textual data.
- To verify the benefits of the progressive generative process, we analyze the ProGBN from an information theory perspective, and reveal it can well alleviate the well-known latent variable collapse issue .
- Experiments on different corpora show that our models outperform other popular NTMs in extracting deeper interpretable topics and deriving better multi-layer document representation.

2. Related work

Deep Topic Models Deep PTMs(Blei et al., 2010; Paisley et al., 2014; Gan et al., 2015; Zhou et al., 2016; Zhao et al., 2018; Zhou et al., 2015; Zhou & Carin, 2013) are developed to constructed multi-layer document representations, with adjacent layers connected through specific factorization. For instance, gamma belief network (GBN)(Zhou et al., 2015) is constructed via factorizing the shape parameters of the gamma distributed latent representations; DPFA(Gan et al., 2015) extends PFA(Zhou & Carin, 2013) into a multi-layer version; DirBN(Zhao et al., 2018) is developed via factorizing the Dirichlet distributed topic matrix. Besides, there are various interests in building effective deep NTMs. For example, (Zhang et al., 2018) proposed a deep VAE framework for deep neural topic modeling with the latent representation followed a Weibull distribution. And (Duan et al., 2021a) explored designing efficient sawtooth structures with the word embedding technique. Further, to address the posterior collapse issue, Li et al. (2022a) propose a policy gradient based training algorithm for deep NTMs.

Knowledge Informed Deep Topic Models Although prior knowledge is incorporated into deep topic models in many works, the methods of incorporation vary. Specifically, JoSH (Meng et al., 2020) adopted a highly effective strategy that utilizes a category hierarchy as guidance and models the semantic correlation between category words



(a) Concept hierarchy (b) Semantic graph $A^{(1)}$ (c) Semantic graph $A^{(2)}$

(d) Bottom level Bow $X^{(1)}$ (e) Middle level Bow $X^{(2)}$ (f) Top level Bow $X^{(3)}$

Figure 1. (a) Concept hierarchy and (b) $\sim$ (c) are the first and second layer semantic graphs built from concept hierarchy, respectively; (d) is the original bag-of-words (Bow) representation for a document; and (e) $\sim$ (f) are the middle and top level Bow built with text coarsening, respectively.

through joint spherical text and tree embedding. The works (Duan et al., 2021b) and (Wang et al., 2022) used pre-defined knowledge graphs as a regularization to instruct models on learning hierarchical topics, but they do not modify the generative model; and HyperMiner (Xu et al., 2022) introduced hyperbolic embeddings to facilitate the mining of implicit semantic hierarchy. Different from the previous works, this paper presents a novel generative model that gradually generates textual data while incorporating prior knowledge via word representations or knowledge graphs.

3. Progressive Generative Process for Bayesian Deep Topic Model

This section will first describe a textual data coarsening process (§ 3.1), followed by the introduction of progressive generative model (§ 3.2) and corresponding embedding based decoder (§ 3.3). After that, we give details for inference network (§ 3.4) and model training algorithm (§ 3.5).

3.1. Knowledge Informed Textual Data Coarsening Process

As shown in Fig. 1(a), words can be organized in a concept hierarchy, where the concepts at adjacent layers following the hypernym relations (Miller, 1995). And a parent concept can be considered to be more abstract in nature than its child concepts. Inspired by the above observation, we build an upward-downward word transfer process for coarsening textual data, in which words first transfer upward to their parent concepts (i.e., more general or abstract concepts) and then randomly transfer downward to children's concepts from parent concepts with the same probability. For example, the word "company" will first transfer to the node "organization", and then randomly transfer to words "company" and "university". This process can be generalized to a criterion where words are randomly transferred to their semantically related words and themselves with the same probability, and naturally implemented by a semantic graph that represent the relationships among words. (The semantic graph of the concept hierarchy in Fig. 1(a) is shown in Fig. 1(b) and Fig. 1(c)). In the following, we will describe a general framework for textural data coarsening process and provide two instances of this general framework.

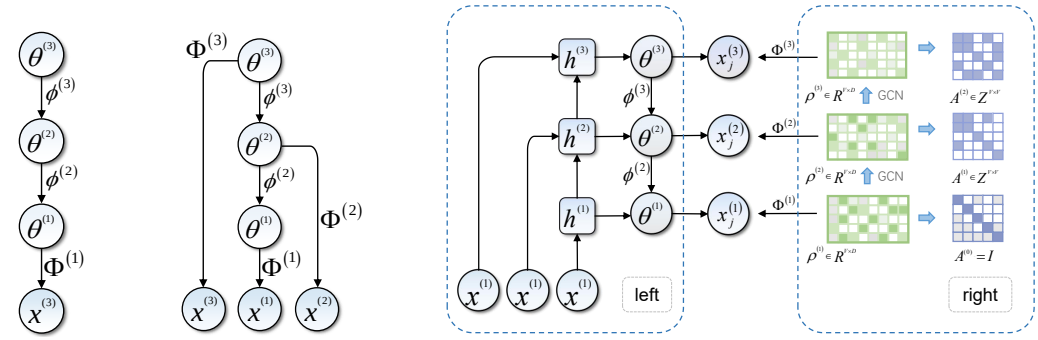
Specifically, given a text corpus consisting of J documents $X = \{\boldsymbol{x}_j\}_{j=1}^J$, the tth token in the jth document can be represented as a one-hot vector $\boldsymbol{x}_{j,t} \in \mathbb{Z}^V$, where V denotes the vocabulary size; and given a hierarchical word semantic graph $\{\mathcal{G}^{(l)}\}_{l=1}^{L-1}$, where the higher level graph represent more general semantic relationship among words, which can be represent as a adjacent matrix $\{A^{(l)} \in \mathbb{Z}^{V\times V}\}_{l=1}^{L-1}$. And the textual data coarsening process of each token $\boldsymbol{x}_{n,t}^{(1)}$ can be defined as:

$$q(\boldsymbol{x}_{j,t}^{(l+1)} \mid \boldsymbol{x}_{j,t}^{(l)}) = \text{Cat}(\boldsymbol{x}_{j,t}^{(l)};\ p = \boldsymbol{x}_{j,t}^{(l-1)} Q^{(l)}) \tag{1}$$

with

$$\{Q_{i,j}^{(l)} = A_{i,j}^{(l)} \,/\, \textstyle\sum_{j=1}^{V} A_{i,j}^{(l)}\}_{i=1,j=1}^{V,V} \tag{2}$$

where, $\text{Cat}(\boldsymbol{x};\ p)$ is a categorical distribution over the one-hot row vector $\boldsymbol{x}$ with probabilities given by the row vector $\boldsymbol{p}$, and $Q^{(l)} \in \mathbb{R}_+^{V\times V}$ is the probability transition matrix that each row sum to one, $\boldsymbol{x}^{(l-1)}Q^{(l)}$ is to be understood as a row vector-matrix product. Under the proposed framework, each document $\boldsymbol{x}_j^{(1)}$ in the corpus can be augmented to a multi-level representation $\{\boldsymbol{x}_j^{(l)}\}_{l=1}^{L}$, where L is the number



(a) GBN (b) Progressive GBN (c) Variational inference network

Figure 2. The graphical model of (a) Gamma belief network (GBN), (b) Progressive GBN; (c) The variational inference network of Progressive GBN, consisting of a hierarchical latent variable inference network (left) and a graph neural network (GCN) based variational embedding inference network (right) .

of levels. In the following, we provide two methods to construct hierarchical word semantic graph.

Knowledge Graph Structure: Given a pre-defined knowledge graph such as WordNet (Miller, 1995), as shown in figure. 1(a), we can construct a hierarchical graph as:

$$\{A^{(l)}\}_{m,n} = 1 \text{ if } \omega_m \text{ and } \omega_n \text{ have a same ancestor at } l\text{th layer else } 0. \tag{3}$$

Token Embedding Distance: Given a pre-trained word embedding, such as Glove (Pennington et al., 2014), we can construct hierarchical graph as:

$$\{A^{(l)}\}_{m,n} = 1 \text{ if } \omega_m \text{ is one of a } T^{(l)}\text{-nearest neighbors of } \omega_n \text{ else } 0, \tag{4}$$

where $T^{(l)} \in \{1, 2, \cdots, \}$ is a hyper-parameters.

Remark 3.1. From the Eq. 1 and Eq. 2 , the textural data coarsening process can be seen as a process of building dependencies among words with pre-defined semantic graphs.

3.2. Progressive Generative Model

To model the data constructed by the coarsening process, we extend GBN (Zhou et al., 2015) to propose a progressive generative model, progressive gamma belief network. Generally, given the multi-level bag-of-words (Bow) representation $\{\boldsymbol{x}_n^{(l)} \in \mathbb{Z}^V\}_{l=1}^L$, as shown in Fig. 2(b), the generative model with L layers can be formulated as

$$\begin{aligned}
&\boldsymbol{\theta}_j^{(L)} \sim \text{Gam}(r,\, c_j^{(L+1)}),\ \boldsymbol{x}_j^{(L)} \sim \text{Pois}(\boldsymbol{\Phi}^{(L)}\boldsymbol{\theta}_j^{(L)}),\\
&\cdots,\\
&\boldsymbol{\theta}_j^{(l)} \sim \text{Gam}(\boldsymbol{\phi}^{(l+1)}\boldsymbol{\theta}_j^{(l+1)},\, c_j^{(l+1)}),\ \boldsymbol{x}_j^{(l)} \sim \text{Pois}(\boldsymbol{\Phi}^{(l)}\boldsymbol{\theta}_j^{(l)}),\\
&\cdots,\\
&\boldsymbol{\theta}_j^{(1)} \sim \text{Gam}(\boldsymbol{\phi}^{(2)}\boldsymbol{\theta}_j^{(2)},\, c_j^{(2)}),\ \boldsymbol{x}_j^{(1)} \sim \text{Pois}(\boldsymbol{\Phi}^{(1)}\boldsymbol{\theta}_j^{(1)}),
\end{aligned} \tag{5}$$

where, $\boldsymbol{\phi}^{(l+1)} \in \mathbb{R}_+^{K^{(l)} \times K^{(l+1)}}$ is the factor loading matrix at layer l; $\boldsymbol{\theta}^{(l)} \in \mathbb{R}_+^{K^{(l)}}$ denotes the gamma distributed latent representation (topic proportions) of layer l; $\boldsymbol{\Phi}^{(l)} \in \mathbb{R}_+^{K^{(l)} \times V}$ can be regarded as the topic matrix for the observation $\boldsymbol{x}_j^{(l)}$ at l layer; $K^{(l)}$ is the number of topic at layer l.

The ProGBN first factorize the count vector $\boldsymbol{x}_j^{(1)}$ (e.g., the bag-of-word of document j as the product of the factor loading matrix $\boldsymbol{\Phi}^{(1)}$ (topics), and gamma distributed factor scores $\boldsymbol{\theta}_j^{(1)}$ (topic proportions), under the Poisson likelihood; for $l \in \{1, \ldots, L-1\}$, the shape parameter of gamma distributed hidden units $\boldsymbol{\theta}_j^{(l)} \in \mathbb{R}_+^{K^{(l)}}$ is further factorized into the product of the connection weight matrix $\boldsymbol{\Phi}^{(l+1)} \in \mathbb{R}_+^{K^{(l)} \times K^{(l+1)}}$ and hidden units $\boldsymbol{\theta}_j^{(l+1)}$ of layer $l+1$, capturing the dependence between different layers; the augment vector $\boldsymbol{x}_j^{(l)}$ is generated by drawing from the Poisson distribution with rate parameter $\boldsymbol{\Phi}^{(l)}\boldsymbol{\theta}_j^{(l)}$; the top layer's hidden units $\boldsymbol{\theta}_j^{(T)}$ share the same $r \in \mathbb{R}_+^{K^{(T)}}$ as their gamma shape parameters; and $c_j^{(t+1)}$ are gamma scale parameters, which set as 1 in our models.

3.3. Embedding-Based Topic Generative Process

For ProGBN, we need to create two class decoders: $\{\boldsymbol{\phi}^{(l)}\}_{l=2}^L$, which captures the relationship between adjacent layer latent variables, and $\{\boldsymbol{\Phi}^{(l)}\}_{l=1}^L$, which captures the relationship between latent variables and words. Motivated from the recent popular distributed topic representation adopted in neural topic models (Dieng et al., 2020; 2019b), we employ the Sawtooth Connector technique (Duan et al., 2021a) to build decoder $\{\boldsymbol{\phi}^{(l)}\}_{l=2}^L$ as:

$$\begin{aligned}
&\boldsymbol{\alpha}_k^{(l)} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}),\ k \in \{1, 2, \cdots, K^{(l)}\},\ l \in \{1, \cdots, L\}\\
&\boldsymbol{\phi}_k^{(l)} = \text{Softmax}({\boldsymbol{\alpha}^{(l-1)}}^T \boldsymbol{\alpha}_k^{(l)}),\ l \in \{2, \cdots, L\}
\end{aligned} \tag{6}$$

where $\boldsymbol{\alpha}_k^{(l)} \in \mathbb{R}^D$ is a distributed semantic representation

of kth topic at layer l and modeled as a Gaussian distributed variables, which aim to model the stochasticity of topic interrelationships and provide a appropriate uncertainty.

To create an effective decoder $\{\mathbf{\Phi}^{(l)}\}_{l=1}^{L}$, one way (Li et al., 2022a) is employing embeddings to capture the relationship between topics and words, which can be defined as:

$$\begin{aligned} &\boldsymbol{\rho}_v \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}), \ v \in \{1, 2, \cdots, V\}, \\ &\mathbf{\Phi}_k^{(l)} = \text{Softmax}(\boldsymbol{\alpha}^{(l)} \boldsymbol{\rho}), \quad l \in \{1, \cdots, L\} \end{aligned} \tag{7}$$

where $\boldsymbol{\rho}_v \in \mathbb{R}^D$ is a distributed representation of the vth words. While this decoder can capture the dependency of topics and words, it ignores the semantic relationships among words that are constructed in the textual data coarsening process, which is discussed in Reamrk. 3.1. Generally, as shown in Eq. 1, the observation at the high level is constructed from the observation at the low level through the semantic relations between words, which is represented by a graph, as shown in Fig. 1(b) and 1(c). Thus, the observations at different levels will contain the semantic relationships represented by the corresponding layer graph. To capture this semantic relationships, we develop a novel graph-enhanced decoder, which can be described as:

$$\begin{aligned} &\boldsymbol{\rho}_v^{(l)} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}), \ v \in \{1, 2, \cdots, V\}, \ l \in \{1, \cdots, L\} \\ &\mathbf{\Phi}_k^{(l)} = \text{Softmax}(\boldsymbol{\alpha}^{(l)} \boldsymbol{\rho}^{(l)}), \quad l \in \{1, \cdots, L\} \\ &\boldsymbol{A}_{v_1 v_2}^{(l)} \sim \text{Bern}(\sigma(\boldsymbol{\rho}_{v_1}^{(l+1)} W \boldsymbol{\rho}_{v_2}^{(l+1)})), \quad l \in \{1, \cdots, L-1\}, \end{aligned} \tag{8}$$

where $\boldsymbol{\rho}_v^{(l)} \in \mathbb{R}^D$ is a distributed representation of the vth words at layer l, $\sigma(\cdot)$is the sigmoid function and $\text{Bern}(\cdot)$ denotes the Bernoulli distribution; W is a learnable parameter. By modeling the semantic graph of different layers with Bernoulli likelihood, we can better incorporate the semantic relationship among words to the corresponding layer decoder (Kipf & Welling, 2016b; Shen et al., 2021).

3.4. Variational Inference Network

The inference network of the proposed model is built around two main components: variational encoder and the variational embedding inference network.

Variational Latent Variables Inference: As shown in Fig. 2(c)(left), we employ a Weibull upward-downward variational encoder to approximate the posteriors of gamma distribution latent variables $\{\boldsymbol{\theta}_j^{(l)}\}_{l=1}^{L}$ as:

$$\begin{aligned} &\boldsymbol{q}(\boldsymbol{\theta}_j^{(l)} \mid \{\boldsymbol{x}_j^{(t)}\}_{t=1}^{l}, \boldsymbol{\theta}_j^{(l+1)}) = \text{Weibull}(\text{NN}_k(\tilde{\boldsymbol{h}}_j^{(l)}), \text{NN}_\lambda(\tilde{\boldsymbol{h}}_j^{(l)})), \\ &\tilde{\boldsymbol{h}}_j^{(l)} = \text{NN}_c(\boldsymbol{h}_j^{(l)}, \boldsymbol{\theta}_j^{(l+1)}), \ \boldsymbol{h}_j^{(l)} = \text{NN}_h(\boldsymbol{x}_j^{(l)}, \ \boldsymbol{h}_j^{(l-1)}), \end{aligned} \tag{9}$$

where $\text{NN}(\cdot)$ are deep neural networks, $\tilde{\boldsymbol{h}}_j^{(l)}$ combine the upward information of document feature $\boldsymbol{h}_j^{(l)}$ and downward information of latent variable $\boldsymbol{\theta}_j^{(l+1)}$. The details of the inference network can be found in Appendix.A.2.

Variational Embedding Inference: Due to the topic embedding $\{\boldsymbol{\alpha}^{(l)}\}_{l=1}^{L}$ and the word embedding at bottom layer $\{\boldsymbol{\rho}^{(1)}\}$ are full data driven parameters, we approximate their posteriors as:

$$\begin{aligned} &q(\boldsymbol{\alpha}_k^{(l)}) = \mathcal{N}(W_{\alpha,k,\mu}^{(l)}, W_{\alpha,k,\sigma}^{(l)}), \ l \in \{1, \cdots, L\}, \\ &q(\boldsymbol{\rho}_v^{(1)}) = \mathcal{N}(W_{\rho,v,\mu}^{(1)}, W_{\rho,v,\sigma}^{(l)}), \end{aligned} \tag{10}$$

where $\{W_{\cdot,\cdot,\cdot}^{(l)}\}$ are learnable parameters. To approximate the posterior of higher layer word embedding $\{\boldsymbol{\rho}^{(l)}\}_{l=2}^{L}$, we need to consider two semantic relationships, where the first is the relationship between word embedding at different layer and the second is the relationship among words introduced by the semantic graph. Fortunately, the Eq. 1 describes the textual data coarsening process with mathematics form, and motivates us build a corresponding word embedding transfer process, which can preserve the relationship among words. In particular, as shown in Fig. 2(c)(right), we construct the variational posterior of word embeddings with graph neural network (Kipf & Welling, 2016a) as:

$$q(\boldsymbol{\rho}_v^{(l+1)} \mid \boldsymbol{\rho}^{(l)}, A^{(l)}) = \mathcal{N}(\tilde{A}_{v:}^{(l)} \boldsymbol{\rho}^{(l)} W_{\mu,v}^{(l)}, \ \tilde{A}_{v:}^{(l)} \boldsymbol{\rho}^{(l)} W_{\sigma,v}^{(l)}) \tag{11}$$

where, $l \in \{1, \cdots, L-1\}$, and $A^{(l)} \in \mathbb{Z}^{V \times V}$ are adjacent matrixes which are built in Sec. 3.1; $\tilde{A}^{(l)} = D^{-\frac{1}{2}} A^{(l)} D^{-\frac{1}{2}}$ is the normalized adjacent matrix with degree matrix D.

3.5. Inference and Estimation

The optimization objective of ProGBN can be achieved by maximizing the evidence lower bound (ELBO) of the log marginal likelihood, which can be computed as:

$$\begin{aligned} \mathcal{L} = &\sum_{j=1}^{J} \sum_{l=1}^{L} \mathbb{E}_Q \left[\ln p(\boldsymbol{x}_j^{(l)} \mid \boldsymbol{\theta}_j^{(l)}, \boldsymbol{\alpha}^{(l)}, \boldsymbol{\rho}^{(l)}) \right] \\ &+ \gamma \sum_{l=1}^{L} \mathbb{E}_Q \left[\ln p(A^{(l)} \mid \boldsymbol{\rho}^{(l)}) \right] \\ &- \sum_{j=1}^{J} \sum_{l=1}^{L} \mathbb{E}_Q \left[\ln \frac{q(\boldsymbol{\theta}_j^{(l)} \mid \boldsymbol{x}_j^{(1)}, \boldsymbol{x}_j^{(l)}, \boldsymbol{\theta}^{(l)})}{p(\boldsymbol{\theta}_j^{(l)} \mid \boldsymbol{\theta}_j^{(l+1)}, \boldsymbol{\alpha}^{(l)}, \boldsymbol{\alpha}^{(l+1)})} \right] \\ &- \sum_{l=1}^{L} \mathbb{E}_Q \left[q(\boldsymbol{\rho}^{(l)}) / p(\boldsymbol{\rho}^{(l)}) \right] - \sum_{l=1}^{L} \mathbb{E}_Q \left[q(\boldsymbol{\alpha}^{(l)}) / p(\boldsymbol{\alpha}^{(l)}) \right] \end{aligned} \tag{12}$$

Where, $Q = q(\boldsymbol{\theta}_j \mid -) \, q(\boldsymbol{\alpha}) \, q(\boldsymbol{\rho})$, and γ denote the hyper-parameter and is set as 0.05 in our experiments. The first two terms are the expected log-likelihood or reconstruction error of corresponding layer observation and semantic graph respectively, while last three terms are the Kullback–Leibler (KL) divergence that constrains variational posterior to be close to its prior in the generative model. The parameters in ProGBN can be directly optimized by advanced gradient algorithms, like Adam (Kingma & Ba, 2014).

4. Analysing ProGBN from the Principle of Information Theory

The "latent variable collapse" is a common problem in hierarchical variational autoencoders (HVAEs) (Sønderby et al., 2016; Dieng et al., 2019a; Maaløe et al., 2019; Li et al., 2022b). As a HAVE, the deep neural topic models, such as WHAI and SawETM, also encounter this issue obviously (Li et al., 2022a). In this section, we employ theory analysis to confirm that the proposed model can well alleviate the "latent variable collapse" issue.

4.1. Analysis: Text Coarsening Process

Rethinking the text coarsening process as defined in Eq. 1, this process can be regarded as a Markov chain. As a result, we can use the data processing inequality (Thomas & Joy, 2006) to obtain the mutual information relationships between the original observation and the higher lever augment observation as follows:

$$I(X^{(1)};\, X^{(1)}) \geq I(X^{(1)};\, X^{(2)}), \cdots, \geq I(X^{(1)};\, X^{(L)}) \quad (13)$$

where $I(\cdot)$ denote the mutual information between two variables, $x^{(l)}$ is the observation at lth layer, and the equal sign is true when the constructed semantic graph $A^{(l)} = I$.

4.2. Analysis: Progressive Generative Process

Assumption 4.1. The encoder are ideal information transmission models which can perfectly preserve all the input information during the inference process.

Under Assumption. 4.1, we analyse ProGBN with the principle of information theory (Thomas & Joy, 2006), and derive a lower bound for the mutual information between the observation $X^{(1)}$ and higher layer latent variables $\boldsymbol{\theta}^{(\geq l)}$. Generally, given the original observation $X^{(1)}$ and corresponding augment observation $\{X^{(l)}\}_{l=2}^{L}$ by the text coarsening process, we have:

$$I(X^{(1)};\, \boldsymbol{\theta}^{(\geq l)}) \geq I(X^{(1)};\, X^{(l)}) \quad (14)$$

where, $\boldsymbol{\theta}^{(\geq l)} = \{\boldsymbol{\theta}^{(l)}, \cdots, \boldsymbol{\theta}^{(L)}\}$ are the latent variables, and $I(\cdot)$ denote the mutual information between two variables, and the detailed derivation can be found in Appendix. A.1.

As VAE-like models, deep NTMs inherit the phenomenon of latent variable collapse from traditional VAEs, where the variational posterior collapses to the prior and provides meaningless latent representations at higher layers. In other words, the hierarchical latent variable model cannot guarantee the mutual information (MI) $I(X^{(1)};\, \boldsymbol{\theta}^{(\geq l)})$ between the observed data and the hidden variables at higher levels, which may decrease to zero as the number of layers increases and is independent of the observed data (Li et al., 2022b). Based on the assumption of our model, we can derive a lower bound for the mutual information $I(X^{(1)};\, \boldsymbol{\theta}^{(\geq l)})$, which highlights the advantege of ProGBN in terms of alleviating "latent variable collapse" issue.

5. Experiment

5.1. Experimental Setup

Datasets: Our experiments are conducted on four widely used benchmark datasets of varying sizes, including *20 News Groups* (20NG)(Lang, 1995), *Tag My News* (TMN) (Vitale et al., 2012), *Reuters* extracted from the Reuters-21578 dataset(R8), *Reuters Corpus Volume I (RCV1)*. In particular, 20NG and TMN, R8 are the three corpora that are associated with document labels. We follow the procedure in SawETM(Duan et al., 2021a) to preprocess these documents to obtain their BoW representations and the statistics of these datasets are presented in Appendix. B.1.

Baselines: As baselines, we choose several exemplary ones from the state-of-the-art topic models, including: 1) **LDA** (Blei et al., 2003), a basic Bayesian topic model; 2) **AVITM** (Srivastava et al., 2017), a NTM which replaces the mixture model in LDA with a product of experts; 3) **ETM** (Dieng et al., 2020), a embedding topic model that marries LDA with word embeddings; 4) **GBN** (Zhou et al., 2015), an extension of LDA with hierarchical latent variables; 5) **WHAI**(Zhang et al., 2018), a Weibull hybrid autoencoding inference model based on GBN(Zhou et al., 2015); 6) **SawETM** (Duan et al., 2021a), which proposes a Sawtooth Connection module to build the dependencies between topics at different layers; 7) **TopicNet**(Duan et al., 2021b), a knowledge-based hierarchical NTM that guides topic discovery through prior semantic graph; 8)**TopicKGA**(Wang et al., 2022), a knowledge-based hierarchical NTM with adaptive semantic graph; 9) *dc*-**ETM** (Li et al., 2022a) a DNTM which apply skip-connection structure in its hierarchical generative model; 10) **ProGBN-kg/wv**, which employs a knowledge graph(Miller, 1995) or pre-trained word embedding (Pennington et al., 2014) to construct a semantic graph in the text data coarsening process; and we set a variant ProGBN-x that directly use the original Bow as the higher level observation. In this case, the semantic graph $A^{(l)} =: I$. Note that the ProGBN-kg/ProGBN-wv will be reduced to the ProGBN-x if there is no knowledge (all the word or entity is not included in these resources) during the textual data coarsening process.

Experiment Setting 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 . 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, dc-ETMs and ProGBN

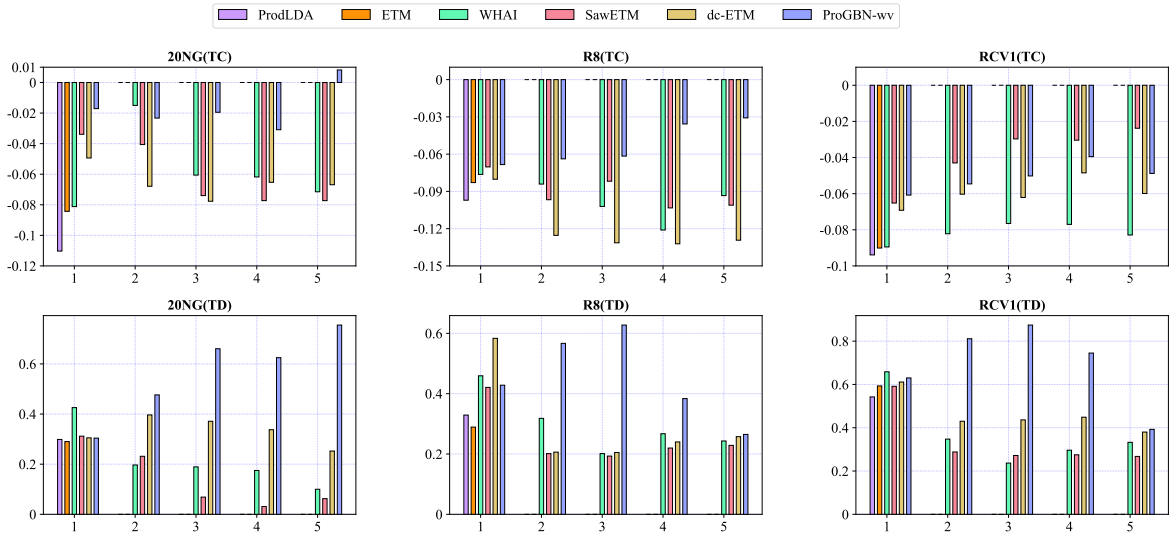


Figure 3. The performance comparison of different models on the topic quality. The top row shows the topic coherence score, i.e., NPMI, and the bottom row displays the topic diversity score. The horizontal axis represents the index of the layers and we set up 5 layers for all the hierarchical topic models.

the mini-batch size as 200. For optimization, we adopt the same Adam optimizer (Kingma & Ba, 2014) with a learning rate of 1e-2.

5.2. Experimental Results

Topic Quality: To make a comprehensive quantitative comparison, we adopt topic coherence(**TC**) and topic diversity (**TD**) to evaluate topic interpretable/quality. Precisely, topic coherence is obtained by taking the average Normalized Pointwise Mutual Information (NPMI) of the top 10 words of each topic (Aletras & Stevenson, 2013). Note that the value of NPMI ranges from -1 to 1, and higher values indicate better interpretability. We use external Wikipedia documents[1] as its reference corpus to estimate the co-occurrence probabilities of words. Following Dieng et al (Dieng et al., 2020), we define Topic diversity to be the percentage of unique words in the top 25 words of all topics. Diversity close to 1 means more diverse topics.

Fig. 3 shows the topic quality comparison results of different models. For the bottom layer topic, the ProGBN gets the best performance for TC on three datasets, which can be attributed to its effective progressive generative process that can provide a good prior. Furthermore, because the latent variables in the bottom layer frequently do not suffer from the latent variable collapse problem, there are no differences in TD performance between different topic models. For the higher layer topics, it's evident that the ProGBN performs best for TD on all the datasets, which can be attributed to its ability to keep the mutual information between the observation and higher layer latent variables, as discussed at Sec.4.

[1] https://github.com/dice-group/Palmetto

While the dc-ETM also gets better TD results compared with SawETM by modeling the skip-connection in its generative process, it can't achieve a better TC compared with SawETM from the bottom layer to the top layer. The reason behind this may be the information in the observed data will be scattered into the various layer variables, leading to the attenuation of information. According to another principle, ProGBN translates information through its forward process, which prevents the information from being scattered. And as a result, ProGBN can achieve the best TC results on the 20NG and R8 datasets while producing comparable TC by-products on the RCV1 dataset with sufficient improvement on TD compared with SawETM.

Document Modeling: To measure the document modeling performance, we use the average of pre-heldout-word **perplexity** (the lower, the better) to measure the performance. Similar to Zhang et al. (2018), for each corpus, we randomly select 80% of the word token from each document to form a training matrix T, holding out the remaining 20% to form a testing matrix Y. The detailed results are shown in Tab. 2. The deep topic model's performance generally outperforms shallow ones, demonstrating the effectiveness of deep structure in improving modeling capability. Benefiting from the progressive generative process, both ProGBN-kg, and ProGBN-wv achieve lower perplexity scores than other deep topic models. And ProGBN-wv gets the best performance, which may be attributed to the pre-trained word embedding containing more semantic information than the knowledge graph. Further, while the ProGBN-x achieves comparable results with other deep topic models, there is a clear gap between the ProGBN-x and ProGBN-kg/wv, which proves the importance of knowledge-informed textu-

Table 1. Document clustering results comparison (km-Purity and km-NMI) on the 1st hidden layer or the concatenation of all hidden layers of different topic models. The best scores of each dataset are highlighted in boldface.

Methods	Layer	km-Purity(%)			km-NMI(%)		
		20NG	TMN	R8	20NG	TMN	R8
LDA	1	41.79 ± 0.75	48.17 ± 0.86	75.74 ± 0.73	45.15 ± 0.89	30.96 ± 0.87	39.82 ± 0.94
AVITM	1	42.33 ± 0.58	55.28 ± 0.40	78.96 ± 0.42	46.33 ± 0.48	35.57 ± 0.38	41.20 ± 0.52
ETM	1	42.61 ± 0.63	59.35± 0.59	80.20 ± 0.43	48.40 ±0.56	38.75 ± 0.80	41.28 ± 0.72
GBN	1	43.30 ± 0.47	50.89 ± 0.93	76.52 ± 0.15	46.51 ± 0.96	31.34 ± 0.72	41.24 ± 0.87
WHAI	1	42.35 ± 0.79	45.06 ± 0.88	75.70 ± 0.81	46.98 ± 1.03	37.34 ± 0.78	43.98 ± 0.94
SawETM	1	43.33 ± 0.64	62.02 ± 0.85	82.25 ± 0.79	50.77 ± 0.75	40.78 ± 0.84	42.97 ± 0.93
TopicNet	1	40.88 ± 0.76	59.80 ± 0.842	78.06 ± 0.73	47.85 ± 0.942	38.06 ± 0.79	40.58 ± 0.85
TopicKGA	1	42.02 ± 0.79	63.48 ± 0.84	80.15 ± 0.97	51.45 ± 0.80	38.54 ± 0.79	41.08 ± 0.81
dc-ETM	1	40.11 ± 0.86	50.12 ± 0.92	71.30 ± 0.67	44.12 ± 0.92	35.02 ± 0.84	38.34 ± 0.78
ProGBN-x	1	47.36 ± 0.74	59.83 ± 0.29	84.29 ± 0.65	50.74 ± 0.23	37.43 ± 0.54	46.91 ± 0.83
ProGBN-kg	1	54.31 ± 0.41	63.52 ± 0.24	84.12 ± 0.67	56.01 ± 0.90	**40.94** ± 0.57	49.36 ± 0.38
ProGBN-wv	1	**54.68** ± 0.44	**64.23** ± 0.57	**84.64** ± 0.74	**57.41** ± 0.78	40.27 ± 0.93	**51.41** ± 0.60
GBN	All	41.17 ± 0.34	47.21 ± 0.84	72.93 ± 0.54	44.20 ± 0.92	30.02 ± 0.57	31.35 ± 0.73
WHAI	All	32.00 ± 0.77	47.21 ± 0.85	70.80 ± 0.68	39.33 ± 0.84	30.02 ± 0.90	41.25 ± 0.87
SawETM	All	38.69 ± 0.86	55.56 ± 0.92	75.89 ± 0.67	39.33 ± 0.92	32.72 ± 0.84	39.55 ± 0.78
dc-ETM	All	48.60 ± 0.84	58.75 ± 0.62	78.29 ± 0.64	55.79 ± 0.93	38.43 ± 0.71	48.62 ± 0.76
ProGBN-x	All	50.24 ± 0.76	61.72 ± 0.34	85.93 ± 0.54	56.75 ± 0.85	39.13 ± 0.64	51.22 ± 0.47
ProGBN-kg	All	**57.56** ± 0.64	68.42 ± 0.32	**87.91** ± 0.28	57.46 ± 0.59	41.31 ± 0.47	53.98 ± 0.85
ProGBN-wv	All	57.14 ± 0.38	**69.47** ± 0.54	87.23 ± 0.82	**58.12** ± 0.46	**41.56** ± 0.78	**54.16** ± 0.58

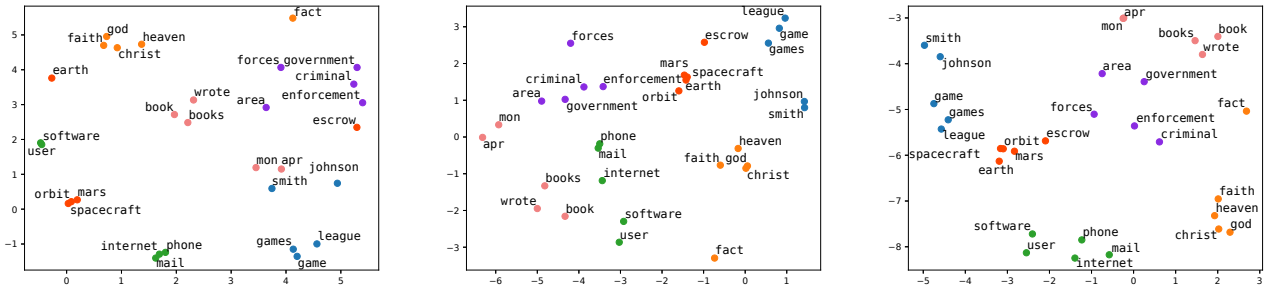
ral data coarsening processes. Another experimental finding is that the degree of improvement of ProGBN-kg/ProGBN-wv over ProGBN-x is different on different datasets. This may be due to the fact that the amount of information obtained from the knowledge graph or word representation may be different for different datasets, which makes the degrees of improvement of ProGBN-kg/ProGBN-wv over ProGBN-x different. Besides, we performed an ablation study on the graph-enhanced decoder, and the experimental results verified its effectiveness.

Document Representation: Since per-document topic proportions can be viewed as unsupervised document representations, we intend to evaluate the quality of such representations by performing document clustering tasks. In detail, we use the trained topic models to extract the latent representations of the testing documents and then apply K-Means to predict the clusters. We use the purity and normalized mutual information metric (NMI) to measure the KMeans clusters (denoted by **km-Purity** and **km-NMI**) (the higher, the better). The clustering results are exhibited in Tab. 1. The results of only using the bottom layer latent feature for clustering demonstrate that ProGBN-x/kg/wv significantly improves the performance compared with other deep topic models. This highlights the effectiveness of the progressive generative process, which can provide a good prior for the bottom layer latent variables. Compared to the results on the TMN dataset, the improvement of ProGBN is more pronounced on the 20NG and R8 datasets. This could be because ProGBN is better suited for modeling complex data, and the TMN dataset contains shorter text. Unlike other deep topic models such as GBN, WHAI, and SawETM, where concatenating hierarchical latent document representations can negatively impact clustering performance, ProGBN has a positive impact. This is likely due to the fact that ProGBN effectively addresses the issue of latent collapse (Li et al., 2022a) and is able to effectively combine information from different levels of observed variables.

Table 2. Comparisons of hold-out perplexity on different benchmarks.

Methods	Depth	Perplexity		
		20NG	RCV1	R8
LDA	1	735	942	996
ProdLDA	1	784	951	561
ETM	1	742	921	985
GBN	5	678	877	657
WHAI	5	726	906	773
SawETM	5	685	873	530
dc-ETM	5	647	801	420
ProGBN-x	5	653	798	436
ProGBN-kg	5	620	753	411
-w/o graph decoder	5	633	784	419
ProGBN-wv	5	**614**	**735**	**408**
-w/o graph decoder	5	632	769	427

Visualization of Different Layer Word Embeddings: The top 10 words from six topics at the 5th layer are selected



(a) $1th$ layer word embedding $\boldsymbol{\rho}^{(1)}$ (b) $3th$ layer word embedding $\boldsymbol{\rho}^{(3)}$ (c) $5th$ layer word embedding $\boldsymbol{\rho}^{(5)}$

Figure 4. T-SNE visualisation of different layer word embeddings. The top 10 words from six topics at the $5th$ layer are selected for visualization. The different color represent different topics at the $5th$ layer

Table 3. The 5th-layer topics learned by SawETM, dc-ETM and ProGBN on 20NG, where each topic is interpreted by its top-10 words. We select semantic-related topics for comparison, highlighting the semantically related words in boldface.

Topic	SawETM	dc-ETM	ProGBN-wv
1	lines subject organization com article just host don writes know	organization lines subject don **article** just **writes** think com like	**wrote** book **article** published read apr books magazine written volume articles
2	lines subject organization com article just host don writes know	lines organization subject article **baseball games game hockey** time nntp	smith **players game** dave league mike johnson games player john
3	lines subject organization com article just host don writes know	organization lines subject article **windows file** com just writes host	rom macintosh interface intel cpu **software** motorola car computers disk
4	lines subject organization com article just host don writes know	lines subject organization article com **university** nntp host posting just	**university** harvard yale stanford smith professor science dept james research
5	lines subject organization com article just host don writes know	organization lines subject **space gov** **nasa** article writes just host	spacecraft earth orbit result large **nasa** planet increase particularly mars

for visualization in Fig. 4(a) to 4(c) using t-SNE visualization (Van der Maaten & Hinton, 2008). The results show that semantically similar words have a stronger clustering effect at the highest layer, while the effect diminishes at lower layers. For instance, as shown in Fig. 4(c) and 4(a), Semantically similar words like 'earth' and'mars' are close in the high-layer word embedding space but far apart in the bottom layer word embedding space. This can be attributed to the fact that the higher layer is used to capture more abstract semantic structures, while the lower layer captures more specific semantic structures.

Visualization of Topics: The learned topics at the 5th layer of different models are shown in Tab. 3, and more results can be found in Appendix. C.2. It can be observed that the topics learned by SawETM are quite similar, which explains why concatenating its hierarchical latent document representations does not improve and even harms the performance on downstream tasks. Compared to dc-ETM, the topics learned by ProGBN-wv exhibit more semantic consistency and diversity, demonstrating the effectiveness of our proposed model in discovering semantically meaningful topics at higher layers. Additionally, the visualization of hierarchical topics can be found in Appendix C.1.

6. Conclusion

In this paper, we propose a novel ProGBN model that incorporates a knowledge-informed textual data coarsening process and a corresponding text data generation process to enhance the capacity of deep topic modeling. Additionally, we propose a graph-enhanced decoder to capture the semantic dependencies established during the textual data coarsening process. Through theoretical analysis based on information theory principles, we demonstrate the ability of our model to avoid the latent variable collapse problem. Furthermore, we conduct practical experiments to validate the effectiveness of our proposed model.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant U21B2006; in part by Shaanxi Youth Innovation Team Project; in part by the Fundamental Research Funds for the Central Universities QTZX23037 and QTZX22160; in part by the 111 Project under Grant B18039; in part by the Fundamental Research Funds for the Central Universities; in part by the Innovation Fund of Xidian University under Grant YJSJ23016;

References

Aletras, N. and Stevenson, M. Evaluating topic coherence using distributional semantics. In *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013)–Long Papers*, pp. 13–22, 2013.

Austin, J., Johnson, D. D., Ho, J., Tarlow, D., and van den Berg, R. Structured denoising diffusion models in discrete state-spaces. *Advances in Neural Information Processing Systems*, 34:17981–17993, 2021.

Bansal, A., Borgnia, E., Chu, H.-M., Li, J. S., Kazemi, H., Huang, F., Goldblum, M., Geiping, J., and Goldstein, T. Cold diffusion: Inverting arbitrary image transforms without noise. *arXiv preprint arXiv:2208.09392*, 2022.

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.

Dieng, A. B., Kim, Y., Rush, A. M., and Blei, D. M. Avoiding latent variable collapse with generative skip models. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 2397–2405. PMLR, 2019a.

Dieng, A. B., Ruiz, F. J., and Blei, D. M. The dynamic embedded topic model. *arXiv preprint arXiv:1907.05545*, 2019b.

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 *ICML 2021: International Conference on Machine Learning*, July 2021a.

Duan, Z., Xu, Y., Chen, B., Wang, C., Zhou, M., et al. Topicnet: Semantic graph-guided topic discovery. *Advances in Neural Information Processing Systems*, 34:547–559, 2021b.

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.

Gu, J., Zhai, S., Zhang, Y., Bautista, M. A., and Susskind, J. f-dm: A multi-stage diffusion model via progressive signal transformation. *arXiv preprint arXiv:2210.04955*, 2022.

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

Karras, T., Aila, T., Laine, S., and Lehtinen, J. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017.

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.

Kipf, T. N. and Welling, M. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016a.

Kipf, T. N. and Welling, M. Variational graph auto-encoders. *arXiv preprint arXiv:1611.07308*, 2016b.

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

Langley, P. Crafting papers on machine learning. In Langley, P. (ed.), *Proceedings of the 17th International Conference on Machine Learning (ICML 2000)*, pp. 1207–1216, Stanford, CA, 2000. Morgan Kaufmann.

Lee, S., Chung, H., Kim, J., and Ye, J. C. Progressive deblurring of diffusion models for coarse-to-fine image synthesis. *arXiv preprint arXiv:2207.11192*, 2022.

Li, Y., Wang, C., Duan, Z., Wang, D., Chen, B., An, B., and Zhou, M. Alleviating "posterior collapse" in deep topic models via policy gradient. In *NeurIPS 2022: Neural Information Processing Systems*, Dec. 2022a.

Li, Y., Wang, C., Xia, X., Liu, T., Miao, X., and An, B. Out-of-distribution detection with an adaptive likelihood ratio on informative hierarchical vae. In *Advances in Neural Information Processing Systems*, 2022b.

Maaløe, L., Fraccaro, M., Liévin, V., and Winther, O. Biva: A very deep hierarchy of latent variables for generative modeling. In *Advances in neural information processing systems*, pp. 6551–6562, 2019.

Marius, M. K. N. C. J. and Burkhardt, K. S. Hierarchical topic evaluation: Statistical vs. neural models.

Meng, Y., Zhang, Y., Huang, J., Zhang, Y., Zhang, C., and Han, J. Hierarchical topic mining via joint spherical tree and text embedding. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, pp. 1908–1917, 2020.

Miao, Y., Grefenstette, E., and Blunsom, P. Discovering discrete latent topics with neural variational inference. In *International Conference on Machine Learning*, pp. 2410–2419. PMLR, 2017.

Miller, G. A. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.

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.

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.

Pennington, J., Socher, R., and Manning, C. D. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 1532–1543, 2014.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Razavi, A., Van den Oord, A., and Vinyals, O. Generating diverse high-fidelity images with vq-vae-2. *Advances in neural information processing systems*, 32, 2019.

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.

Shen, D., Celikyilmaz, A., Zhang, Y., Chen, L., Wang, X., Gao, J., and Carin, L. Towards generating long and coherent text with multi-level latent variable models. *arXiv preprint arXiv:1902.00154*, 2019.

Shen, D., Qin, C., Wang, C., Dong, Z., Zhu, H., and Xiong, H. Topic modeling revisited: A document graph-based neural network perspective. *Advances in Neural Information Processing Systems*, 34, 2021.

Shu, R. and Ermon, S. Bit prioritization in variational autoencoders via progressive coding. In *International Conference on Machine Learning*, pp. 20141–20155. PMLR, 2022.

Sønderby, C. K., Raiko, T., Maaløe, L., Sønderby, S. K., and Winther, O. Ladder variational autoencoders. In *Advances in neural information processing systems*, pp. 3738–3746, 2016.

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

Tan, B., Yang, Z., AI-Shedivat, M., Xing, E. P., and Hu, Z. Progressive generation of long text with pretrained language models. *arXiv preprint arXiv:2006.15720*, 2020.

Thomas, M. and Joy, A. T. *Elements of information theory*. Wiley-Interscience, 2006.

Van der Maaten, L. and Hinton, G. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

Vitale, D., Ferragina, P., and Scaiella, U. Classification of short texts by deploying topical annotations. In *European Conference on Information Retrieval*, pp. 376–387. Springer, 2012.

Wang, D., Xu, Y., Li, M., Duan, Z., Wang, C., Chen, B., Zhou, M., et al. Knowledge-aware bayesian deep topic model. *Advances in Neural Information Processing Systems*, 35:14331–14344, 2022.

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.

Xiao, Y., Wu, L., Guo, J., Li, J., Zhang, M., Qin, T., and Liu, T.-y. A survey on non-autoregressive generation for neural machine translation and beyond. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

Xu, Y., Wang, D., Chen, B., Lu, R., Duan, Z., and Zhou, M. Hyperminer: Topic taxonomy mining with hyperbolic embedding. In Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., and Oh, A. (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 31557–31570. Curran Associates, Inc., 2022.

Zhang, H., Chen, B., Guo, D., and Zhou, M. WHAI: Weibull hybrid autoencoding inference for deep topic modeling. In *International Conference on Learning Representations*, 2018.

Zhang, H., Chen, B., Cong, Y., Guo, D., Liu, H., and Zhou, M. Deep autoencoding topic model with scalable hybrid Bayesian inference. *To appear in IEEE TPAMI*, 2020. URL http://arxiv.org/abs/2006.08804.

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

Zhou, M. and Carin, L. Negative binomial process count and mixture modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(2):307–320, 2013.

Zhou, M., Hannah, L., Dunson, D., and Carin, L. Beta-negative binomial process and Poisson factor analysis. In *Artificial Intelligence and Statistics*, pp. 1462–1471, 2012.

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.

A. Model

A.1. Detailed Derivations for Section 3.1

Assumption A.1. The encoder are ideal information transmission models that can perfectly preserve all the input information during the inference process.

Under Assumption. A.1, we have

$$\mathcal{H}\left(\boldsymbol{\theta}^{(\geq l)} \mid X^{(l)}\right) = 0 \tag{15}$$

We define the joint information entropy for the latent variables $(X^{(1)}, X^{(l)}, \boldsymbol{\theta}^{(\geq l)})$ as

$$\begin{aligned} &\mathcal{H}\left(X^{(1)}, X^{(l)}, \boldsymbol{\theta}^{(\geq l)}\right) \\ &= \mathcal{H}\left(X^{(l)}, \boldsymbol{\theta}^{(\geq l)}\right) + \mathcal{H}\left(X^{(1)} \mid X^{(l)}, \boldsymbol{\theta}^{(\geq l)}\right) \\ &= \mathcal{H}\left(X^{(l)}\right) + \mathcal{H}\left(\boldsymbol{\theta}^{(\geq l)} \mid X^{(l)}\right) + \mathcal{H}\left(X^{(1)} \mid X^{(l)}, \boldsymbol{\theta}^{(\geq l)}\right) \end{aligned} \tag{16}$$

we define the joint information entropy for the latent variables $(X^{(1)}, X^{(l)})$ as

$$\mathcal{H}\left(X^{(1)}, X^{(l)}\right) = \mathcal{H}\left(X^{(l)}\right) + \mathcal{H}\left(X^{(1)} \mid X^{(l)}\right) \tag{17}$$

According to the chain rule of information entropy, we can get

$$\mathcal{H}\left(X^{(1)}, X^{(l)}, \boldsymbol{\theta}^{(\geq l)}\right) \geq \mathcal{H}\left(X^{(1)}, X^{(l)}\right) \tag{18}$$

The Eq. 18 is equal when $\mathcal{H}(\boldsymbol{\theta}^{(\geq l)} \mid X^{(1)}, X^{(l)}) = 0$. According to Eq. 15 $\sim$ 18, we can get

$$\therefore \mathcal{H}\left(X^{(l)}\right) + \mathcal{H}\left(\boldsymbol{\theta}^{(\geq l)} \mid X^{(l)}\right) + \mathcal{H}\left(X^{(1)} \mid X^{(l)}, \boldsymbol{\theta}^{(\geq l)}\right) = \mathcal{H}\left(X^{(l)}\right) + \mathcal{H}\left(X^{(1)} \mid X^{(l)}\right) \tag{19}$$

Which can further get

$$\mathcal{H}\left(X^{(1)} \mid X^{(l)}, \theta^{(\geq l)}\right) = \mathcal{H}\left(X^{(1)} \mid X^{(l)}\right) \tag{20}$$

Similarity, we can get

$$\mathcal{H}(X^{(1)} \mid \boldsymbol{\theta}^{(\geq l)}) \geq \mathcal{H}(X^{(1)} \mid X^{(l)}, \boldsymbol{\theta}^{(\geq l)}) \tag{21}$$

The Eq. 22 is equal when $\mathcal{H}(X^{(l)} \mid \boldsymbol{\theta}^{(\geq l)}) = 0$.

we define the mutual information between the original observation $X^{(1)}$ and latent variables $\boldsymbol{\theta}^{(\geq l)}$ as

$$\begin{aligned} &I\left(X^{(1)}; \boldsymbol{\theta}^{(\geq l)}\right) = \mathcal{H}\left(X^{(1)}\right) - \mathcal{H}\left(X^{(1)} \mid \boldsymbol{\theta}^{(\geq l)}\right) \\ &\geq \mathcal{H}\left(X^{(1)}\right) - \mathcal{H}(X^{(1)} \mid X^{(l)}, \boldsymbol{\theta}^{(\geq l)}) \text{ (According to Eq. 21)} \end{aligned} \tag{22}$$

we define the mutual information between the original observation $X^{(1)}$ and higher layer observation $X^{(l)}$ as

$$\begin{aligned} &I\left(X^{(1)}; X^{(l)}\right) = \mathcal{H}\left(X^{(1)}\right) - \mathcal{H}\left(X^{(1)} \mid X^{(l)}\right) \\ &= \mathcal{H}\left(X^{(1)}\right) - H\left(X^{(1)} \mid X^{(l)}, \boldsymbol{\theta}^{(\geq l)}\right) \text{ (According to Eq. (20)} \\ &\leq I\left(X^{(1)}; \boldsymbol{\theta}^{(\geq l)}\right) \text{ (According to Eq. (22)} \end{aligned} \tag{23}$$

Thus, we can get

$$\therefore I\left(X^{(1)}; \boldsymbol{\theta}^{(\geq l)}\right) \geq I\left(X^{(1)}; X^{(l)}\right), \tag{24}$$

A.2. Detail of hierarchical variational inference network

To approcimate the gamma distribution latent variables, we utilize a Weibull upward-downward variational encoder to approximate the posteriors of $\{\boldsymbol{\theta}_j^{(l)}\}_{l=1}^L$ inspired by the work in Zhang et al. (2018; 2020). Then we have

$$\boldsymbol{q}(\boldsymbol{\theta}_j^{(l)} \mid \{\boldsymbol{x}_j^{(1)}\}_{l=1}^l, \boldsymbol{\theta}_j^{(l+1)}) = \text{Weibull}(\boldsymbol{k}_j^{(l)}, \boldsymbol{\lambda}_j^{(l)}), \tag{25}$$

where parameters $\boldsymbol{k}_j^{(l)}, \boldsymbol{\lambda}_j^{(l)} \in \mathbb{R}_+^{K_l}$ are deterministic transformations of the observed document features $\boldsymbol{x}_j^{(1)}$, corresponding level observation $\boldsymbol{x}_j^{(l)}$, and the information from the stochastic up-down path $\boldsymbol{\theta}_j^{(l+1)}$. Fig. **??** shows how these pieces of information are propagated to influence $\boldsymbol{\theta}_j^{(l+1)}$. Formally, the inference process can be described by

$$\begin{aligned}
\boldsymbol{h}_j^{(l)} &= \text{Re}LU(\boldsymbol{h}_j^{(l-1)}\boldsymbol{W}_1^{(l)} + \boldsymbol{b}_1^{(l)}),\\
\tilde{\boldsymbol{h}}_j^{(l)} &= \begin{cases} \boldsymbol{h}_j^{(L)} \oplus (\boldsymbol{x}_j^{(L)}\boldsymbol{W}_2^{(l)} + \boldsymbol{b}_2^{(l)})), \ l = L, \\ \boldsymbol{h}_j^{(l)} \oplus \boldsymbol{\Phi}_i^{(l+1)}\boldsymbol{\theta}_j^{(l+1)} \oplus (\boldsymbol{x}_j^{(l)}\boldsymbol{W}_2^{(l)} + \boldsymbol{b}_2^{(l)}), \ l < L, \end{cases}\\
\boldsymbol{k}_j^{(l)} &= \ln\,(1 + \exp(\tilde{\boldsymbol{h}}_j^{(l)}\boldsymbol{W}_3^{(l)} + \boldsymbol{b}_3^{(l)})),\\
\boldsymbol{\lambda}_j^{(l)} &= \ln\,(1 + \exp(\tilde{\boldsymbol{h}}_j^{(l)}\boldsymbol{W}_4^{(l)} + \boldsymbol{b}_3^{(l)})),
\end{aligned} \tag{26}$$

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

A.3. Detain of marginal likelihood

For PG-GGN, with the multi-level observation $\{\boldsymbol{x}_j^{(l)}\}_{l=1}^L$, the marginal likelihood of the dataset X is defined as

$$\begin{aligned}
&\boldsymbol{p}(\{X, \boldsymbol{A}^{(l)}\}_{l=1}^L \mid \{\boldsymbol{\alpha}^{(l)}, \boldsymbol{\rho}^{(l)}, \boldsymbol{W}^{(l)}\}_{l=1}^L)\\
&= \int\int\int \prod_{l=1}^{L}\prod_{j=1}^{J} \boldsymbol{p}(\boldsymbol{x}_j^{(l)} \mid \boldsymbol{\alpha}^{(l)}, \boldsymbol{\rho}^{(l)}) \prod_{l=1}^{L} \boldsymbol{p}(\boldsymbol{A}^{(l)} \mid \boldsymbol{\rho}^{(l)})\\
&\quad \prod_{l=1}^{L}\prod_{j=1}^{J} \boldsymbol{p}(\boldsymbol{\theta}_j^{(l)} \mid \boldsymbol{\theta}_j^{(l+1)}, \boldsymbol{\alpha}^{(l)}, \boldsymbol{\alpha}^{(l+1)})\\
&\quad \prod_{l=1}^{L} \boldsymbol{p}(\boldsymbol{\alpha}^{(l)}) \boldsymbol{p}(\boldsymbol{\rho}^{(l)}) d\boldsymbol{\theta}_{l=1,j=1}^{L,J} d\boldsymbol{\alpha}_{l=1}^{L} d\boldsymbol{\rho}_{l=1}^{L}.
\end{aligned} \tag{27}$$

A.4. Training algorithm

We summarize the training algorithm at Algorithm 1.

B. Experimental detail

B.1. Detaset

Details about the datasets this paper relied on are as follows:

C. Additional results

C.1. Hierarchical topic visualization

Fig .5 displays the hierarchical topic structure with the top two layers and the bottom two layers.

C.2. The learned topics at 5th layer

Algorithm 1 Training algorithm of ProGBN

Set mini-batch size m and the number of layer L
Construct hierarchical word semantic graph $\{\mathrm{A}^{(\mathrm{l})}\}_{\mathrm{l}=1}^{\mathrm{L}-1}$
Generate multi-level representations $\{\mathrm{X}^{(\mathrm{l})}\}_{\mathrm{l}=2}^{\mathrm{L}}$ in a coarsening process according to Eq. (1)
Initialize the encoder parameters $\boldsymbol{\Omega}$ and decoder parameters $\boldsymbol{\Psi}$;
for iter = 1,2, $\cdots$ **do**
  Randomly select a mini-batch of m documents with their multi-level representation to form a subset $\{\mathrm{X}^{(\mathrm{l})}\}_{\mathrm{l}=1}^{\mathrm{L}} = \left\{\boldsymbol{x}_i^{(l)}\right\}_{l=1,m}^{L}$;
  Dram random noise $\left\{\boldsymbol{\varepsilon}_i^l\right\}_{i=1,l=1}^{m,L}$ from uniform distribution;
  Calculate $\nabla_{\boldsymbol{\Omega},\boldsymbol{\Psi}} L\left(\boldsymbol{\Omega},\boldsymbol{\Psi};\{\mathrm{X}^{(\mathrm{l})}\}_{\mathrm{l}=1}^{\mathrm{L}},\{\mathrm{A}^{(\mathrm{l})}\}_{\mathrm{l}=1}^{\mathrm{L}},\left\{\boldsymbol{\varepsilon}_i^l\right\}_{i=1,l=1}^{m,L}\right)$ according to Eq. (12), and update encoder parameters $\boldsymbol{\Omega}$ and decoder parameter $\boldsymbol{\Psi}$ jointly ;
end for

Table 4. Statistics of the datasets.(N: Dataset size. L: Average document length. V: Vocabulary size. C: Number of categories.)

Dataset	Dataset size	Vocabulary size	Average document length.	Number of categories
20NG	18,846	2,000	108	20
TMN	32,597	13,368	18	7
R8	7,639	2,074	47	8
RCV2	804,414	8,000	74	N/A

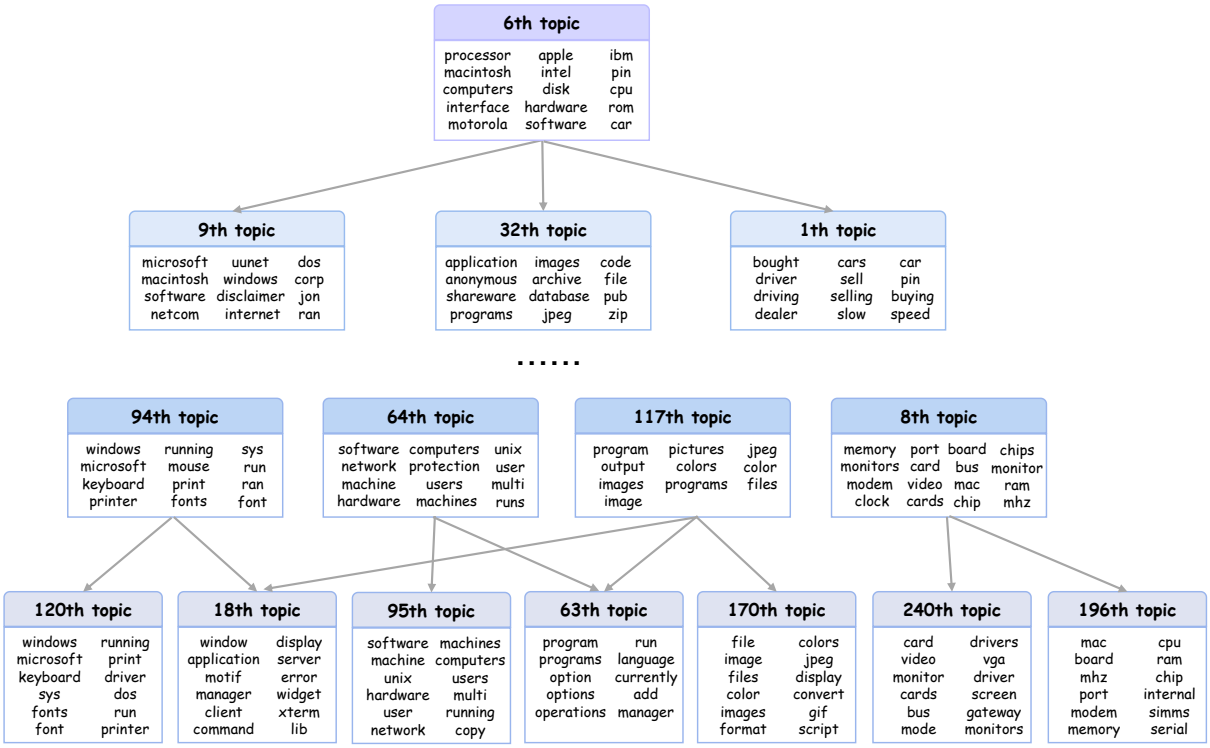


Figure 5. An example of hierarchical topics learned from 20NG by a 5-layer ProGBN-wv, We only show example topics at the top two layers and bottom two layers.

Table 5. The 5th-layer topics learned by SawETM and ProGBN-wv on 20NG, where each topic is interpreted by its top-10 words.

Topic	SawETM	ProGBN
1	lines subject organization com article just host don writes know	window error include lib jpeg function application widget display code
2	lines subject organization com article just host don writes know	key encryption clipper attack chip secure security government public keys
3	lines subject organization com article just host don writes know	lines com subject organization article host posting just university nntp
4	lines subject organization com article just host don writes know	said people went time house came did know day didn
5	lines subject organization com article just host don writes know	religion christian god faith people believe does say evidence christians
6	lines subject organization com article just host don writes know	gun weapon weapons fbi waco batf guns koresh firearms law
7	lines subject organization com article just host don writes know	available image ftp data graphics faq pub file images version
8	lines subject organization com article just host don writes know	medical health disease cancer drugs study age patients drug aids
9	lines subject organization com article just host don writes know	team game games season win hockey baseball play nhl year
10	lines subject organization com article just host don writes know	conference national april american report information new york year professor
11	lines subject organization com article just host don writes know	science theory light energy physics scientific space surface field ray
12	lines subject organization com article just host don writes know	max armenians turkish armenian argic armenia soviet serdar genocide azerbaijan
13	lines subject organization com article just host don writes know	drive scsi bit lines subject card dos organization windows uses
14	lines subject organization com article just host don writes know	israel israeli jews jewish arab arabs muslims islam gay homosexual
15	lines subject organization com article just host don writes know	god jesus christ lord bible shall sin man faith life
16	lines subject organization com article just host don writes know	don just people know think like time going good say