Bidirectional Convolutional Poisson Gamma Dynamical Systems

Wenchao Chen, Chaojie Wang, Bo Chen, Yicheng Liu, Hao Zhang
National Laboratory of Radar Signal Processing
Xidian University, Xi’an, Shaanxi 710071, China
wcchen_xidian@163.com, xd_silly@163.com, bchen@mail.xidian.edu.cn,
moooooore66@gmail.com, zhanghao_xidian@163.com

Mingyuan Zhou
McCombs School of Business
The University of Texas at Austin
Austin, TX 78712, USA
mingyuan.Zhou@mccombs.utexas.edu

Abstract

Incorporating the natural document-sentence-word structure into hierarchical Bayesian modeling, we propose convolutional Poisson gamma dynamical systems (PGDS) that introduce not only word-level probabilistic convolutions, but also sentence-level stochastic temporal transitions. With word-level convolutions capturing phrase-level topics and sentence-level transitions capturing how the topic usages evolve over consecutive sentences, we aggregate the topic proportions of all sentences of a document as its feature representation. To consider not only forward but also backward sentence-level information transmissions, we further develop a bidirectional convolutional PGDS to incorporate the full contextual information to represent each sentence. For efficient inference, we construct a convolutional-recurrent inference network, which provides both sentence-level and document-level representations, and introduce a hybrid Bayesian inference scheme combining stochastic-gradient MCMC and amortized variational inference. Experimental results on a variety of document corpora demonstrate that the proposed models can extract expressive multi-level latent representations, including interpretable phrase-level topics and sentence-level temporal transitions as well as discriminative document-level features, achieving state-of-the-art document categorization performance while being memory and computation efficient.

1 Introduction

How to represent documents to capture their underlying semantic structures is a key research problem in text analysis and language modeling [1-4]. It has been a long-standing challenge to capture long-range dependency in sequential data, especially for the word sequences of long documents that take discrete values for a large vocabulary. A simple remedy that helps capture long-range word dependency, but at the expense of sacrificing local structure, is to simplify the representation of each document as a bag of words (BoW) that ignores word order. Under this simplified BoW text representation, probabilistic topic models, such as latent Dirichlet allocation (LDA) [5,6], Poisson

*Equal Contribution.
†Corresponding Author.

factor analysis (PFA) and their various generalizations, have been widely deployed for text analysis, providing semantically meaningful latent representations. Vanilla topic models usually assume that each word is generated from a topic, characterized by a specific distribution over the terms in the vocabulary. Extracting a set of latent topics from a corpus, they characterize each document with its proportion over these topics. To enhance their modeling power and interpretability of vanilla topics models, which often have a single stochastic layer, a variety of deep generalizations with multiple stochastic layers have been proposed. Despite providing improvement, their ultimate potentials have been limited by their use of BoW that completely ignores word local structure.

To remedy the loss of word order, Wang et al. propose convolutional PFA (CPFA) as a convolutional generalization of PFA, extracting phrase-level (i.e., n-gram) rather than word-level topics, where the number of words in each phrase is determined by the convolutional filter size that is typically set as \( n = 3 \). However, CPFA still ignores sentence order, limiting its power in capturing longer-range dependency beyond the filter size.

Treating text as word sequences, deep neural networks, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs), are wildly used to learn text representations. To efficiently capture long-distance relationships and obtain more expressive representations, there is a recent trend to construct a hierarchical language model that models each document as a sequence of sentences, and each sentence as a sequence of words, achieving promising results in learning good document representations for downstream tasks. For example, Tang et al. employ CNNs/RNNs to extract sentence representations, and summarize them into a document representation with another temporal architecture that models the sentence-level transitions. Yang et al. introduce attention mechanism into the hierarchical language structure of Tang et al. to further improve its classification performance. Moreover, the widely used Bidirectional Encoder Representations from Transformers (BERT) has been extended in a hierarchical fashion to hierarchical BERT (HIBERT) for document summarization. However, many of these deep neural network based methods need pretrain with a large amount of extra text data, followed by downstream task-specific finetune, to achieve good performance, and it is still often a challenge for these black-box methods to visualize and explain the semantic meanings learned by them.

Moving beyond the constraints of previous work, we represent each document as a sequence of sentences, each sentence as a sequence of word tokens, and each word token as a one-hit vector of dimension \( V \), where \( V \) is the vocabulary size. This provides a lossless representation of a corpus that respects the natural semantic structure of each document. Under such a lossless representation, we first propose convolutional Poisson gamma dynamical system (conv-PGDS) that adopts probabilistic convolutional structure to extract phrase-level features and models the inter-sentence dependency by a dynamical network. To exploit the sentence order, we further employ a bidirectional setting on our model and propose bidirectional conv-PGDS (bi-conv-PGDS). For scalable training and fast out-of-sample prediction, we integrate a stochastic gradient MCMC and a convolutional-recurrent variational inference network to perform posterior inference. We further propose an attention bi-conv-PGDS (attn-bi-conv-PGDS) for supervised learning, combining the representation power of bi-conv-PGDS, discriminative power of deep neural networks, and selection power of the attention mechanism under a principled probabilistic framework. On a variety of text corpora, we show the proposed models can not only capture long-distance relationships by exploiting the natural hierarchical structure inside each document, but also inherit various virtues of probabilistic topic models. The main contributions of this work are summarized as follows:

- Leveraging both word and sentence order, we propose conv-PGDS and its bidirectional generation, bi-conv-PGDS, to incorporate the document/sentence/word hierarchical structure into Bayesian hierarchical modeling. They capture both intra-sentence structure with convolution and inter-sentence dependency with gamma Markov chains. To the best of our knowledge, bi-conv-PGDS is the first unsupervised bidirectional hierarchical probabilistic model for document modeling.
- To achieve scalable training and fast testing, we perform Bayesian inference by combining stochastic gradient-MCMC (SG-MCMC) and convolutional-recurrent variational inference networks.
- We develop a supervised version of bi-conv-PGDS for document categorization and incorporate attention mechanism into it to further enhance its performance.
- Without requiring expensive pretrain on huge extra data, our models achieve state-of-the-art results, with low memory requirement and fast computation, in a variety of document modeling tasks.
2 Convolutional dynamical systems for document modeling

We first present conv-PGDS that integrates the convolutional components of CPFA [20] to model intra-sentence structure, dynamic components of PGDS [43] to model inter-sentence dependency, and a probabilistic pooling operation, which handles sentences of different lengths, into a well-defined generative model for document analysis. We then add a bidirectional generalization. To help better understand the hierarchical model, we summarize the notations in Table 4 in Appendix.

2.1 Convolutional Poisson gamma dynamical systems

Denote $D_j = \{X_{j1}, \ldots, X_{jt}\}$ as the $j$th document with $T_j$ sentences and $X_{jt} = \{x_{jt1}, \ldots, x_{jtl_{jt}}\}$ as its $t$th sentence of length $L_{jt}$, where $x_{jtl}$ is the $l$th word token in this sentence represented with a one-hot vector as $x_{jtl} = (x_{jt1l}, \ldots, x_{jtl})' \in \{0, 1\}^V$, with $x_{jtvl} = 1$ if and only if token $l$ matches term $v$ of the vocabulary of size $V$. Distinct from CPFA that considers word order but ignores sentence order, we introduce conv-PGDS that not only captures the internal temporal structures of each sentence using convolution, but also models the information transmissions between consecutive sentences using a gamma Markov chain. The generative process of conv-PGDS is expressed as

$$w_{jtk} \sim \text{Gam}(\tau_0 v_k, 1/\tau_0), \quad \text{for } k = 1, 2, \ldots, K, \quad w_{jtk} \sim \text{Gam} \left(\tau_0 (\Pi_{jk} - \sum_{s=1}^L w_{jtl-1:s} S_{jt-1}) 1_{S_{j1} = 1}/\tau_0 \right),$$

$$X_{jt} = 1 (M_{jt} > 0), \quad M_{jt} \sim \text{Pois} \left(\delta_{jt} \sum_{k=1}^K D_k * w_{jtk}, D_k (:) \sim \text{Dir} (\eta 1_{V^F})\right),$$

as described in detail below. Conv-PGDS factorizes each sentence $X_{jt} = \{x_{jt1}, \ldots, x_{jtl_{jt}}\}$ as $\{0, 1\}^V \times L_{jt}$, represented with $L_{jt}$ sequential one-hot vectors, under the Bernoulli-Poisson link [43]. It defines $X_{jt}$ by thresholding count matrix $M_{jt} \in \{0, 1, 2, \ldots\}^V \times L_{jt}$ that is factorized under the Poisson likelihood. It uses $\delta_{jt} > 0$ as a sentence-specific scaling factor and $D_k = (d_{k1}, \ldots, d_{kF}) \in \mathbb{R}_+^{V \times F}$ as the $k$th convolutional filter (i.e., $F$-gram phrase-level topic), whose filter size is $F$ and vectorized form is $D_k (:) = (d_{k1}, \ldots, d_{kF})' \in \mathbb{R}_+^{VF}$. A Dirichlet prior is used to impose a simplex constraint on $D_k (:)$. We denote the convolutional representation of sentence $t$ in document $j$ as $W_{jt} = (w_{jtl1}, \ldots, w_{jtlK})' \in \mathbb{R}_+^{K \times S_{jt}}$, where $w_{jtlk} \in \mathbb{R}_+^{S_{jt}}$ is the feature map of sentence $t$ under topic $k$, and $S_{jt} := L_{jt} - F + 1$ varies with the sentence length $L_{jt}$. To capture the semantic or syntactic relations between the sentences in a document, which is often a challenging task [31], we assume $w_{jtl}$ to be gamma distributed and factorize its shape parameters using transition matrix $\Pi \in \mathbb{R}_+^{K \times K}$, which captures cross-sentence temporal dependence, and the average pooling of the feature representation of sentence $t - 1$, expressed as $\sum_{s=1}^{S_{jt-1}} w_{jtl-1:s} S_{jt-1}$. We denote $\Pi_{jk}, \text{ as the } k$th row of $\Pi$ and $w_{jtl-1:s} \text{ as the } s$th column of $W_{jt-1}$. Distinct from traditional deterministic feedforward pooling operations, which cut off the backward message passing, the probabilistic pooling layers in conv-PGDS can be trained jointly with the whole network, which enables the inter-sentence information transmissions to influence $\{w_{jt}\}_T$, the intra-sentence convolutional representations. Inspired by related temporal construction [43], which uses a (truncated) gamma process and the interactions of its atom weights to define a well-regularized transition matrix with tractable inference, we introduce $K$ factor weights $\nu = (v_1, \ldots, v_K)$ to model the strength of each component and let

$$\pi_k \sim \text{Dir} (v_1 \nu_k, \ldots, v_K \nu_k), \quad v_k \sim \text{Gam} (\gamma_0/K, 1/\beta), \quad \text{for } k = 1, 2, \ldots, K,$$

where $\pi_k = (\pi_{1k}, \ldots, \pi_{Kk})'$ is the $k$th column of $\Pi$ and $\pi_{k1:k}$ can be interpreted as the strength of transitioning from topic $k$ to topic $k$. To complete the model, we let $\delta_{jt}, \xi, \beta \sim \text{Gam}(\epsilon_0, 1/\epsilon_0)$.

Similar to CPFA [20], conv-PGDS extracts both global cooccurrence patterns and local temporal structures by forming $F$-gram phrase-level topics into $D_k$, such as “how do you” and “microsoft email address” when $F = 3$. Distinct from CPFA, it extracts $\Pi$ to model temporal dependence across sentences, capturing the transition patterns between the topics exhibited by consecutive sentences (e.g., if “how do you” is an active topic in a sentence, then it is likely to activate its next sentence to exhibit topic “I am fine”). In conv-PGDS, embedded into the sentence-level feature $w_{jt}$ are both phrase-level information of sentence $t$ and transition information of sentences $1$ to $t - 1$.

To summarize, conv-PGDS models both word order, by convolving over sequentially ordered one-hot word vectors, and sentence order, by adding dependency between the latent features of consecutive sentences. Note conv-PGDS reduces to CPFA [20] if ignoring its gamma Markov chain that models sentence order, and to PGDS [43] if removing both the convolution and pooling operations.
2.2 Bidirectional convolutional Poisson gamma dynamical systems

We expand conv-PGDS in (1) to bidirectional conv-PGDS (bi-conv-PGDS), as shown in Fig. 1(b), which is a bidirectional dynamical model trained using all available input information in both the past and future of a specific time frame. The generative model of bi-conv-PGDS is expressed as

\[
\bar{w}_{j_1k} \sim \text{Gam} \left( \tau_0 \bar{\Pi}_k, \sum_{s=1}^{S_{j,t+1}} \bar{w}_{j_1+1,s} / S_{j,t+1}, 1 / \tau_0 \right), \quad \bar{w}_{j_{t+1}k} \sim \text{Gam} \left( \tau_0 \bar{v}_k, 1 / \tau_0 \right),
\]

\[
w_{j_1k} \sim \text{Gam} \left( \tau_0 \Pi_k, \sum_{s=1}^{S_{j,t+1}} w_{j_{t+1},s} / S_{j,t+1}, 1 / \tau_0 \right), \quad w_{j_{t+1}k} \sim \text{Gam} \left( \tau_0 v_k, 1 / \tau_0 \right),
\]

\[
X_{jt} = 1 (M_{jt} > 0), \quad M_{jt} \sim \text{Pois} \left( \delta_t \left( \sum_{k=1}^{K} D_k * \left( \bar{w}_{jtk} + \bar{w}_{jtk} \right) \right) \right), \quad D_k (\cdot) \sim \text{Dir} (\eta_{VF}),
\]

where \( \bar{w}_{jt} \) is the forward hidden feature map and \( w_{jt} \) is the backward one. We let \( u_{jt} = \bar{w}_{jt} + w_{jt} \) to summarize the neighboring sentences around sentence \( t \). To interpret the bidirectional temporal relationship of the topics \( D_k \) in (3), we list the expected value of \( M_{jt} \) as

\[
E \left[ M_{jt} | D, \bar{w}_{j_{t-1}}, \bar{w}_{j_{t+1}}, \bar{\Pi}, \bar{\Pi} \right] = D * \left[ \bar{\Pi} \left( \bar{w}_{j_{t-1}, / S_{j_{t-1}}} 1_{S_{j_{t-1}}} \right) \right] + D * \left[ \bar{\Pi} \left( \bar{w}_{j_{t+1}, / S_{j_{t+1}}} 1_{S_{j_{t+1}}} \right) \right],
\]

which shows that due to the bidirectional dynamic structure, the expected value of \( M_{jt} \) is divided into two parts: one from the forward feature representation, while the other from the backward one.

In addition, \( \bar{\Pi} \) and \( \Pi \) play the role of transiting the latent representations across sentences. Distinct from conv-PGDS, \( u_{jt} \) inferred by bi-conv-PGDS contains transition information of not only past sentences \( 1 \rightarrow t - 1 \), but also future ones \( t + 1 \leftrightarrow T_j \).

3 Bayesian inference and extensions to supervised learning

In this section, we first introduce a Gibbs sampling algorithm for bi-conv-PGDS and a corresponding stochastic gradient MCMC algorithm for scalable inference. To make our models both fast in out-of-sample prediction and easy to incorporate extra side information like document labels, we further construct a novel Weibull distribution based convolutional-recurrent variational inference network. In addition, we introduce a supervised learning component enhanced with attention modules, which can be jointly trained with bi-conv-PGDS using a hybrid SG-MCMC and variational inference.

3.1 Gibbs sampling

By generalizing the variable augmentation and marginalization techniques related to the Poisson, gamma, categorical, and Dirichlet distributions [2][3], we propose a Gibbs sampling algorithm for bi-conv-PGDS, with all update equations and their derivations deferred to the Appendix. Here we note that the local variables of documents can be updated in parallel in each iteration, and the variable augmentation operations for sentence representation learning can also be parallelized at the sentence level. Thus the time cost of each Gibbs sampling iteration can be greatly reduced with Graphical Process Units (GPUs). While the computation of the Gibbs sampler can be accelerated inside each iteration, it requires processing all documents in each iteration and hence has limited scalability.
For scalable inference, we update our global parameters by generalizing TLASGR-MCMC [17, 18], a SG-MCMC algorithm that is developed to sample simplex-constrained global parameters in a mini-batch learning setting and provide high sampling efficiency by preconditioning its gradient with the Fisher information matrix. We defer the details to the Appendix.

3.2 Hybrid SG-MCMC and convolutional-recurrent autoencoding variational inference

Regardless of whether Gibbs sampling or TLASGR-MCMC is used, the need to perform a sampling based iterative procedure at the testing time limits the efficiency of the model for out-of-sample prediction. For the current model and inference, it is also difficult to incorporate the label information. As in Fig. [1(a)] to make our model fast in out-of-sample prediction, we develop a convolutional-recurrent variational inference network to map the observation directly to its latent representations. Following Zhang et al. [18], the Weibull distribution is used to approximate the gamma distributed conditional posterior of $\tilde{w}_{jt}$, $\tilde{w}_{jt}$. As illustrated in Fig. [1(b)], we first introduce an autoencoding variational distribution as $q(\tilde{w}_{jt}) = \text{Weibull}(j,t)$, specifically

$$q(\tilde{w}_{jt}|\cdot) = \text{Weibull}(\hat{\Sigma}_{jt}, \hat{\Pi}_{jt})$$

where $\hat{\Sigma}_{jt}, \hat{\Pi}_{jt} \in \mathbb{R}^{K \times S_j}$ are the parameters of $q(\tilde{w}_{jt})$. To exploit both intra-sentence phrase-level structure and inter-sentence temporal dependencies, we generalize related constructions in Guo et al. [27, 45] to develop a convolutional-recurrent variational inference network, as shown in Fig. [1(b)]. The parameters above are deterministically transformed from the observation $X_{jt}$ as

$$H_{jt} = \text{relu}(C_1 \ast X_{jt} + b_1), \hat{h}_{jt} = \text{GRU}((H_{jt}), (\hat{h}_{jt}), (\hat{h}_{jt})), \hat{\Sigma}_{jt} = \text{exp}(\hat{C}_2 \ast \text{pad}(H_{jt} + \hat{h}_{jt}) + b_2), \hat{\Pi}_{jt} = \text{exp}(\hat{C}_2 \ast \text{pad}(H_{jt} + \hat{h}_{jt} + \hat{b}_2))$$

where $b_1, b_2, \hat{b}_2, \hat{b}_3, \hat{b}_3 \in \mathbb{R}^K$, $C_1 \in \mathbb{R}^{K \times V \times F}$, $\hat{C}_2, \hat{C}_3 \in \mathbb{R}^{K \times K \times F}$. $H_{jt} \in \mathbb{R}^{K \times S_j}$, pad(·) denotes zero-padding, and pool($H_{jt}$) $\in \mathbb{R}^K$ denotes sentence-level average-pooling as defined before. We denote GRU and GRU are the forward and backward part of the bidirectional gated recurrent unit (GRU) [46] and $H_{jt}, \hat{h}_{jt} \in \mathbb{R}^K$. With $H_{jt} + \hat{h}_{jt}$ and $H_{jt} + \hat{h}_{jt}$, realize by first broadcasting $\hat{h}_{jt}$ and $\hat{h}_{jt}$ to match the dimension of $H_{jt}$, we fuse the intra-sentence structural information and inter-sentence temporal dependencies, enabling the proposed convolutional-recurrent variational inference network to learn rich latent representations for bi-conv-PGDS.

To train bi-conv-PGDS, we develop a hybrid SG-MCMC/autoencoding variational inference algorithm by combining TLASGR-MCMC and our proposed convolutional-recurrent variational inference network. Specifically, the global parameters \{$D_{k}\}_{k=1}^{K}$, $\bar{\Pi}$, $\bar{\Pi}$ will be sampled with TLASGR-MCMC, while the parameters of the convolutional-recurrent variational inference network, denoted by $\Omega$, will be updated via stochastic gradient descent (SGD) by maximizing the evidence lower bound (ELBO) [47, 48]. The details are deferred to the Appendix.

3.3 Supervised attention bidirectional convolutional PGDS

To handle the document categorization task, we incorporate label supervision to extend bi-conv-PGDS into supervised bi-conv-PGDS (su-bi-conv-PGDS). To further improve its categorization performance, we combine it with the attention mechanism to develop supervised attention bi-conv-PGDS ( attn-bi-conv-PGDS). Specifically, similar to Long et al. [49], to extract phrases important to the meaning of a sentence, we apply channel-wise attention in our latent feature representation of each sentence as

$$u_{jts} = \text{tanh}(W_u w_{jts} + b_u), \alpha_{jts} = \frac{\exp(w_{jts}^T u_w)}{\sum_i \exp(w_{jts}^T u_w)}, s_{jt} = \sum_s \alpha_{jts} w_{jts}$$
Table 1: Comparison of classification accuracy on unsupervisedly extracted features and runtime per iteration.

<table>
<thead>
<tr>
<th>Methods</th>
<th>Size</th>
<th>Reuters</th>
<th>ELEC</th>
<th>IMDB-2</th>
<th>IMDB-10</th>
<th>Time in seconds</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Accuracy</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LDA [5]</td>
<td>1000</td>
<td>69.2±1.8</td>
<td>75.7±1.8</td>
<td>78.5±1.7</td>
<td>28.6±1.3</td>
<td>18.06</td>
</tr>
<tr>
<td>DocNADE [50]</td>
<td>1000</td>
<td>74.6±0.9</td>
<td>80.4±1.2</td>
<td>81.2±1.3</td>
<td>30.0±0.7</td>
<td>10.17</td>
</tr>
<tr>
<td>DPFA [12]</td>
<td>1000</td>
<td>71.1±1.8</td>
<td>77.2±1.9</td>
<td>78.8±1.9</td>
<td>28.6±1.0</td>
<td>40.34</td>
</tr>
<tr>
<td></td>
<td>1000-500</td>
<td>71.1±1.7</td>
<td>77.8±1.8</td>
<td>79.0±1.7</td>
<td>28.9±0.9</td>
<td>41.68</td>
</tr>
<tr>
<td>DPFA [12]</td>
<td>1000-500-200</td>
<td>71.2±1.7</td>
<td>78.4±1.6</td>
<td>79.3±1.7</td>
<td>29.0±0.9</td>
<td>43.19</td>
</tr>
<tr>
<td>PGBN [13]</td>
<td>1000</td>
<td>72.0±1.4</td>
<td>78.5±1.3</td>
<td>79.7±1.6</td>
<td>29.2±0.8</td>
<td>19.89</td>
</tr>
<tr>
<td></td>
<td>1000-500</td>
<td>73.7±1.3</td>
<td>80.1±1.5</td>
<td>80.9±1.4</td>
<td>30.1±0.6</td>
<td>29.84</td>
</tr>
<tr>
<td>PGBN [13]</td>
<td>1000-500-200</td>
<td>74.8±1.2</td>
<td>80.9±1.3</td>
<td>81.2±1.4</td>
<td>31.0±0.6</td>
<td>33.81</td>
</tr>
<tr>
<td>WHAI [18]</td>
<td>1000</td>
<td>71.2±1.6</td>
<td>78.0±1.7</td>
<td>78.9±1.7</td>
<td>28.9±0.9</td>
<td>9.61</td>
</tr>
<tr>
<td></td>
<td>1000-500</td>
<td>72.7±1.5</td>
<td>79.5±1.6</td>
<td>79.8±1.5</td>
<td>29.7±0.9</td>
<td>12.23</td>
</tr>
<tr>
<td>WHAI [18]</td>
<td>1000-500-200</td>
<td>73.9±1.3</td>
<td>80.1±1.6</td>
<td>80.4±1.4</td>
<td>30.4±1.0</td>
<td>14.12</td>
</tr>
<tr>
<td>DPGDS [51]</td>
<td>1000</td>
<td>73.6±1.0</td>
<td>79.2±0.9</td>
<td>80.2±0.8</td>
<td>31.2±0.6</td>
<td>26.34</td>
</tr>
<tr>
<td>DPGDS [51]</td>
<td>1000-500</td>
<td>74.5±1.0</td>
<td>79.9±0.8</td>
<td>81.0±0.8</td>
<td>32.0±0.6</td>
<td>39.62</td>
</tr>
<tr>
<td>DPGDS [51]</td>
<td>1000-500-200</td>
<td>75.4±0.9</td>
<td>81.4±0.8</td>
<td>81.7±0.8</td>
<td>32.5±0.5</td>
<td>47.75</td>
</tr>
<tr>
<td>CPGBN [20]</td>
<td>1000</td>
<td>75.6±0.8</td>
<td>81.4±0.9</td>
<td>81.8±0.9</td>
<td>34.2±0.5</td>
<td>20.10</td>
</tr>
<tr>
<td>CPGBN [20]</td>
<td>1000-500</td>
<td>76.9±0.7</td>
<td>82.1±0.9</td>
<td>82.3±0.8</td>
<td>35.0±0.6</td>
<td>39.86</td>
</tr>
<tr>
<td>CPGBN [20]</td>
<td>1000-500-200</td>
<td>77.3±0.7</td>
<td>82.5±0.8</td>
<td>82.6±0.7</td>
<td>35.3±0.5</td>
<td>47.54</td>
</tr>
<tr>
<td>conv-PGDS [51]</td>
<td>1000</td>
<td>77.4±0.8</td>
<td>83.1±0.9</td>
<td>83.3±0.9</td>
<td>37.2±0.4</td>
<td>27.50</td>
</tr>
<tr>
<td>conv-PGDS [51]</td>
<td>1000-500</td>
<td>78.0±0.7</td>
<td>84.5±0.8</td>
<td>84.0±0.8</td>
<td>37.9±0.4</td>
<td>29.99</td>
</tr>
</tbody>
</table>

To reward the sentences that provide important clues to correctly classify a document, we again use attention mechanism to measure the importance of different sentences. This yields

$$u_{jt} = \tanh (W_s s_{jt} + b_s), \quad \alpha_{jt} = \frac{\exp(u^T_{jt} u_s)}{\sum_c \exp(u^T_{jt} u_s)}, \quad t_j = \sum_t \alpha_{jt} s_{jt}, \quad (7)$$

where $t_j$ is a document-level feature vector that summarizes all the sentences of a document. Setting the label probability vector as $p_j = (p_{j1}, \ldots, p_{jC}) = \text{softmax}(W_c t_j + b_c)$, we adopt the cross-entropy loss between the ground truth label distribution $p^g_j$ and predicted label distribution $p_j$ as the training loss: $L_g = -\sum_{j=1}^M \sum_{c=1}^C p_{jc}^g \cdot \log (p_{jc})$. Thus the loss function of the entire framework is modified as $L = -L_g + \lambda L_s$, where $L_g$ refers to the negative ELBO of the generative model, and $\lambda > 0$ is a regularization hyper-parameter to balance data generation and label supervision and we set $\lambda = 0.1$ in our experiment.

4 Experiments

We evaluate the effectiveness of our model on five large corpora, including ELEC, IMDB-2, Reuters, Yelp 2014, and IMDB-10, which are described in detail in the Appendix. We fix the hyperparameters of our models as $\tau_0 = 1, \epsilon_0 = 0.1, \gamma_0 = 0.1, \eta = 0.05$ for all experiments. Python (PyTorch) code is provided at https://github.com/BoChenGroup/BCPGDS

4.1 Unsupervised models

We first evaluate various models by extracting document representations in an unsupervised manner and training a linear classifier on these unsupervisedly extracted document features. We compare to a wide variety of topic models, including LDA [5], DocNADE [50], DPFA [12], PGBN [13], WHAI [18], DPGDS [51], and CPGBN [20], where only CPGBN takes one-hot word sequences as its input, while all the others directly operate on BoW vectors that ignore word order. To make a fair comparison, we adopt comparable multilayer structures across all methods and set the same convolutional filter width $F = 3$ for both CPGBN and our models.

We list in Table 2 the network structures of various methods and their corresponding results, which are either quoted from the original papers or reproduced with the code provided by the authors. Given the same network structure, models with BoW input are shown to underperform those with one-hot vector input, suggesting the importance of the convolutional operation that helps leverage the word order information. Conv-PGDS and bi-conv-PGDS both provide clear improvement over CPGBN, suggesting the importance of capturing inter-sentence dependencies. Incorporating the contextual information into sentence representations that are then aggregated, bi-conv-PGDS extracts more expressive document-level features, as suggested by its consistent improvement over conv-PGDS and all baseline methods. For methodological complexity, we also report the average run time of each iteration.
for all methods. Handling each document parallely and accelerating Gibbs sampling procedure with GPU, our models outperform other non-dynamic models with a comparable, or even less, computational burden. In addition to quantitative evaluation, we also visualize the inferred convolutional filters and transition matrices of conv-PGDS, showing the advantages of the proposed Bayesian models over existing “black-box” deep neural networks in terms of having interpretable latent structures. As shown in Fig. 3, from the full transition matrix $\Pi$ learned from ELEC, we visualize the transitions between the 15 most frequently used convolutional filters. We visualize some example filters and how they transit to each other in Fig. 2. We provide in the Appendix the visualizations of each of these 15 filters. For each column of convolutional filter $D_k \in \mathbb{R}^{1 \times 3}$, the top three terms that have the highest weights and their corresponding weights are exhibited. It is particularly interesting to notice that the top terms at different columns can be combined into a variety of interpretable phrases with related semantics. Taking the 11th filter as an example, its corresponding phrase-level topic and transition relationships are visualized in Fig. 2 (a), which shows that filter 11 on “I bought this/an” is more likely to activate a transition to filter 5 on “I thought/used/ordered it”, filter 7 on “it’s a”, and filter 14 on “have found problem”. In addition, as shown in Fig. 2 (b), filter 20 on “this/these product/thing is/range” is the phrase describing “product” and it is more likely to activate a transition to filters 4, 6, 30 that are mainly about “product evaluations”. The transition links shown in Fig. 2 (a) and (c) correspond to column 11 and row 2 in Fig. 3, respectively. Capturing phrase-level topics and inter-sentence information transitions, the document representations of our models are embedded with rich semantic information, making them well suited for downstream tasks.

To verify the efficiency of our generative model in capturing the word order information, we estimate the likelihood of a sentence with shuffled word order. Fig. 4 (left) shows the likelihood decreases as the shuffling rate increases, indicating CPGDS provides a higher confidence on real sentences than orderless ones. Moreover, we further estimate the likelihood of document with shuffled sentence order and observe similar behaviors in Fig. 4 (right), which illustrate that our generative model can also capture the sentence order information in a document.

### 4.2 Supervised models

To evaluate the effectiveness of attn-bi-conv-PGDS, which incorporates the label information and attention mechanism into bi-conv-PGDS, we compare it with various supervised methods on four popular benchmarks and list the results in Table 2. For comparison, we include BoW-based methods: sAV-ITM [52], MedLDA [10], sWHAI [18]; bag-of-n-gram models [53]: SVM-unigrams, SVM-bigrams, SVM-trigrams; CNN/RNN based methods: SVM-wv [54], LSTM-wv [55], CNN-wv [22], CNN-onehot [56]; methods leveraging the document-sentence-word hierarchical structure: Conv-GRNN [31], LSTM-GRNN [31], HAN-AVE, HAN-ATT [57]; and convolutional probabilistic model: sCPGBN.
Table 2: Comparison of classification accuracy and number of model parameters between supervised models

<table>
<thead>
<tr>
<th>Methods</th>
<th>Model size</th>
<th>ELEC</th>
<th>IMDB-2</th>
<th>IMDB-10</th>
<th>Yelp'14</th>
<th>#Param</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM-unigrams</td>
<td>200</td>
<td>83.7</td>
<td>84.9</td>
<td>30.1</td>
<td>50.4</td>
<td>4.10M</td>
</tr>
<tr>
<td>SVM-bigrams</td>
<td>200</td>
<td>84.6</td>
<td>85.7</td>
<td>30.7</td>
<td>51.0</td>
<td>2.10M</td>
</tr>
<tr>
<td>SVM-trigrams</td>
<td>200</td>
<td>86.8</td>
<td>87.2</td>
<td>52.0</td>
<td>52.7</td>
<td>4.10M</td>
</tr>
<tr>
<td>SVM-one-hot</td>
<td>200-100</td>
<td>87.5</td>
<td>88.0</td>
<td>33.5</td>
<td>53.4</td>
<td>4.14M</td>
</tr>
<tr>
<td>SVM-one-hot</td>
<td>200-100-50</td>
<td>87.8</td>
<td>88.2</td>
<td>34.2</td>
<td>54.3</td>
<td>4.15M</td>
</tr>
<tr>
<td>SVMone-hot</td>
<td>-</td>
<td>86.3</td>
<td>87.7</td>
<td>39.9</td>
<td>58.9</td>
<td>-</td>
</tr>
<tr>
<td>SVM-one-hot</td>
<td>-</td>
<td>87.2</td>
<td>88.2</td>
<td>40.9</td>
<td>57.6</td>
<td>-</td>
</tr>
<tr>
<td>SVM-one-hot</td>
<td>-</td>
<td>87.4</td>
<td>88.5</td>
<td>41.3</td>
<td>59.8</td>
<td>-</td>
</tr>
<tr>
<td>SVM-one-hot</td>
<td>-</td>
<td>87.9</td>
<td>86.5</td>
<td>42.4</td>
<td>59.6</td>
<td>-</td>
</tr>
<tr>
<td>SVM-one-hot</td>
<td>-</td>
<td>88.3</td>
<td>89.0</td>
<td>43.6</td>
<td>62.5</td>
<td>5.30M</td>
</tr>
<tr>
<td>SVM-one-hot</td>
<td>-</td>
<td>88.6</td>
<td>89.5</td>
<td>42.5</td>
<td>59.7</td>
<td>4.94M</td>
</tr>
<tr>
<td>SVM-one-hot</td>
<td>-</td>
<td>91.3</td>
<td>91.6</td>
<td>43.2</td>
<td>62.4</td>
<td>6.09M</td>
</tr>
<tr>
<td>LSTM-one-hot</td>
<td>-</td>
<td>91.6</td>
<td>91.8</td>
<td>44.1</td>
<td>63.2</td>
<td>8.10M</td>
</tr>
<tr>
<td>Conv-GRNN</td>
<td>200-100</td>
<td>91.6</td>
<td>91.5</td>
<td>44.5</td>
<td>63.7</td>
<td>4.94M</td>
</tr>
<tr>
<td>HAN-ATT</td>
<td>200-100</td>
<td>91.5</td>
<td>91.6</td>
<td>47.8</td>
<td>67.0</td>
<td>5.56M</td>
</tr>
<tr>
<td>HAN-ATT</td>
<td>200-100</td>
<td>91.7</td>
<td>91.8</td>
<td>49.4</td>
<td>68.2</td>
<td>5.68M</td>
</tr>
<tr>
<td>DocBERT</td>
<td>-</td>
<td>93.4</td>
<td>94.1</td>
<td>54.2</td>
<td>72.0</td>
<td>11.00M</td>
</tr>
<tr>
<td>DocBERT</td>
<td>-</td>
<td>93.4</td>
<td>94.1</td>
<td>54.2</td>
<td>72.0</td>
<td>11.00M</td>
</tr>
<tr>
<td>DocBERT</td>
<td>-</td>
<td>93.4</td>
<td>94.1</td>
<td>54.2</td>
<td>72.0</td>
<td>11.00M</td>
</tr>
<tr>
<td>doca-conv-PGDS</td>
<td>200</td>
<td>91.9±0.2</td>
<td>92.0±0.2</td>
<td>48.9±0.2</td>
<td>68.0±0.3</td>
<td>12.20M</td>
</tr>
<tr>
<td>attn-bi-conv-PGDS</td>
<td>200</td>
<td>92.5±0.2</td>
<td>92.6±0.2</td>
<td>51.2±0.2</td>
<td>70.5±0.3</td>
<td>12.24M</td>
</tr>
<tr>
<td>attn-bi-conv-PGDS</td>
<td>200</td>
<td>93.0±0.3</td>
<td>93.4±0.3</td>
<td>53.8±0.2</td>
<td>71.2±0.3</td>
<td>12.28M</td>
</tr>
</tbody>
</table>

Figure 4: left: likelihood of shuffled sentence; right: likelihood of shuffled document

and Transformer based pre-trained language models: DocBERT [57]. As shown in Table 2, BoW-based methods in general underperform the methods that take word order into consideration. Traditional methods based on hand-crafted features like SVM-unigrams/bigrams/trigrams typically underperform CNN/RNN based methods, such as CNN-wv, LSTM-wv, that learn local dependence structure from the data. CNN/RNN based methods, which do not exploit the hierarchical document structure, clearly underperform the methods that exploit these structures to capture long-term temporal dependence and cross-sentence information transmission, especially for large-scale datasets with many categories. Pretrained on huge extra text data, DocBERT with large memory requirement and high computational complexity recently achieves state-of-the-art document classification performance. By incorporating hierarchical document structure into probabilistic generative models and the attention mechanism, attn-bi-conv-PGDS achieves comparable results with DocBERT but with significantly fewer parameters (about 90% reduction) and faster test speed (more than 92% reduction in testing time, as shown in Table 2). Moreover, in order to validate that our model is able to attend to informative sentences and phrases in a document given the context, we provide two example visualizations of the hierarchical attention of our model and compare them with CPFA equipped with attention in Fig. 5. We select three most prominent sentences in a document for visualization. Blue and red denote the phrase weight and sentence weight, respectively. Darker color represents greater weight (the sentence weight is listed on its left-side). Due to the hierarchical structure, we normalize the word weight by the sentence weight to make sure that only important words in important sentence are emphasized. As we can see, attn-CPFA can only select the phrases carrying strong sentiments, but cannot deal with complex across-sentence context. Taking the document in Fig. 5(b) as an example, attn-CPFA only focuses on phrase “the worst movie”, thus to classify it into
This movie is the worst of the current Star Wars movies. However, I have seen some comments here saying it is the worst movie EVER. That is certainly NOT the case.

5 Conclusion

Respecting the natural hierarchical structure of each document (words form a sentence, sentences form a documents) and incorporating that information into hierarchical Bayesian modeling, we propose convolutional Poisson gamma dynamical systems (conv-PGDS) to capture not only phrase-level topics, by introducing word-level convolutions, but also how the topic usages evolve over consecutive sentences, by modeling sentence-level transitions. To consider both forward and backward sentence-level information transmissions, we further develop a bidirectional conv-PGDS. For scalable inference, we develop a hybrid Bayesian inference algorithm that integrates SG-MCMC and a convolutional-recurrent variational inference network to approximate the posterior. In addition, we incorporate attention mechanism into our model and develop an attention guided bidirectional conv-PGDS for document categorization. Experiments on unsupervised feature extraction and supervised semi-supervised document categorization tasks show that our models can not only extract discriminative document representations, achieving state-of-the-art results with low memory and computational cost, but also inherit the virtues of probabilistic topic models that provide interpretable phrase-level topics and sentence level-temporal transitions.
Broader Impact

The proposed models are probabilistic topic models considering hierarchical temporal information within each document, which can not only achieve state-of-the-art performance in some text analysis tasks, such as unsupervised feature extraction and (semi-)supervised document categorization, but also obtain semantically meaningful topics and latent features as well as interpretable transition relations. Thus, they may be used to detect harmful articles which may contain fake news, violent content, and fraudulent materials. In addition, they can be used to recommend articles with specific contents to users according to their needs. More importantly, as they are able to provide interpretable latent features, one could try to understand why a certain categorization and recommendation has been made by the proposed model for a given article, so more appropriate actions can be taken rather than purely trusting the model itself to make the right decisions.

Although big pre-trained language models, such as BERT [35], GPT2 [58], and GPT3 [59], could be fine-tuned for a variety of natural language processing tasks to achieve state-of-the-art performance in many different settings, they are consuming significant amounts of computing resources, leading to higher energy consumption and CO2 emissions. Comparing with these pretrained big models, the proposed probabilistic models provide customized probabilistic solutions to specific problems, achieving comparable results in the specific task of document categorization but with much lower memory and computational cost, which is beneficial for energy saving and environmental protection.

Providing good performance while maintaining interpretability becomes an even more urgent issue today given the recent trend in building larger and more complex black-box models trained with bigger data, which work well but make it become increasingly more difficult to understand how and why they work well. We hope our work can motivate machine learners to pay more attention to the study of interpretable and compact models. When evaluating the model, we should analyze the model more and find what can help different areas of society. At present, many people tend to strongly emphasize on the numerical performance and pay less attention to the energy cost of training and deploying these models. Meanwhile, the black box feature of deep learning means that the model may fail without clear explanations, so it is difficult to deploy them to applications with high-level safety and stability requirements. An interpretable model, on the other hand, enables people to understand what the model really learns and how it makes decisions, so as to better evaluate the model and understand how and where to deploy it for the benefits of the society.

Acknowledgements

The authors thank the anonymous reviewers for their constructive comments and suggestions that have helped to improve the paper. B. Chen acknowledges the support of the Oversea Talent by Chinese Central Government, the 111 Project (No. B18039), NSFC (61771361), Shanxi Innovation Team Project, and the Innovation Fund of Xidian University. M. Zhou acknowledges the support of the U.S. National Science Foundation under Grant IIS-1812699.

References


