Sawtooth Factorial Topic Embeddings Guided Gamma Belief Network

Zhibin Duan*, Dongsheng Wang*, Bo Chen,

Chaojie Wang, Wenchao Chen, Yewen Li, Jie Ren, Mingyuan Zhou

Xidian University, The University of Texas at Austin, * Equal contribution

Motivation

Document Representation

- Bag-of-words (Bow)
  - ✓ Term-document frequency count matrix
  - ✓ Simple and powerful
  - ※ Extremely sparse matrices

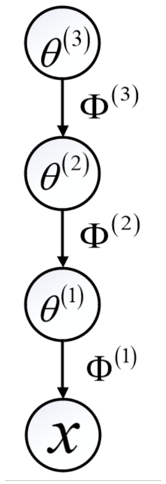
"Yesterday, I bought an earphone. I love it. It is the best electronic product I have bought" ⟹

$$\begin{array}{c} I \\ it \\ love \\ \ldots \\ have \end{array} \quad \begin{bmatrix} 3 \\ 2 \\ 1 \\ \ldots \\ 1 \end{bmatrix}$$

Document Bow

Hierarchical Topic Models

- Gamma belief network (GBN)
  - ✓ Hierarchical document representation
  - ✓ The latent variables at each layer are dependent
  - ※ The topics at each layer are independent



$$\theta_j^{(L)} \sim \mathrm{Gam}\left(r, 1/c_j^{(L+1)}\right),$$

$$\theta_j^{(l)} \sim \mathrm{Gam}\left(\Phi^{(l+1)}\theta^{(l+1)}, 1/c_j^{(l+1)}\right), \quad l = 1, 2, \ldots, L-1$$

$$\phi_k^{(l)} \sim \mathrm{Dir}\left(\eta^{(l)}, \ldots, \eta^{(l)}\right), l = 1, 2, \ldots, L$$

$$x_j \sim \mathrm{Pois}\left(\Phi^{(1)}\theta^{(1)}\right)$$

Graphic Model Generative model

Our contribution

■ Decoder: Sawtooth Factorial Topic Embeddings Guided GBN

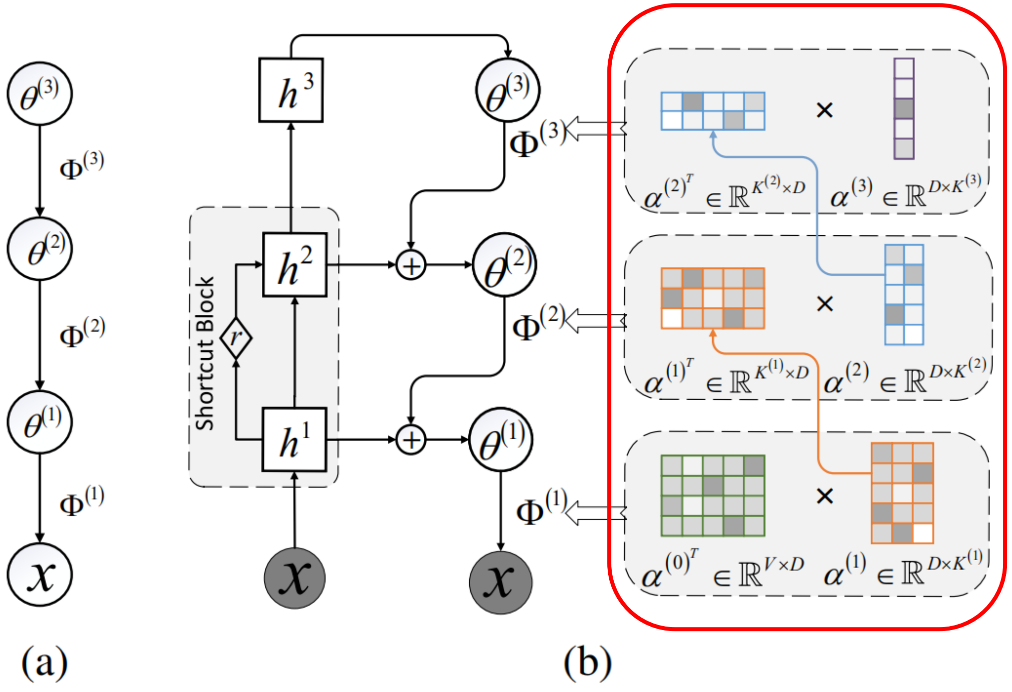


Figure 1. (a) Gamma belief network and (b) overview of the proposed SawETM and its corresponding hierarchical upward and downward encoder networks

$$
\begin{aligned}
&\boldsymbol{\theta}_j^{(L)} \sim \mathrm{Gam}(r, c_j^{(L+1)}),\\
&\boldsymbol{\theta}_j^{(l)} \sim \mathrm{Gam}(\boldsymbol{\Phi}^{(l+1)}\boldsymbol{\theta}_j^{(l+1)}, c_j^{(l+1)}), \quad l = 1, \cdots, L-1\\
&\boldsymbol{\Phi}_k^{(l)} = \mathrm{Softmax}(\boldsymbol{\alpha}^{(l-1)^T}\boldsymbol{\alpha}_k^{(l)}), \quad l = 1, \cdots, L\\
&\boldsymbol{x}_j \sim \mathrm{Pois}(\boldsymbol{\Phi}^{(1)}\boldsymbol{\theta}_j^{(1)}) \qquad (1)
\end{aligned}
$$

- ✓ Preserve latent variables dependency
- ✓ Capture topics dependency
- ✓ Take advantages of the sparsity
- ✓ Project topics in a shared embedding space

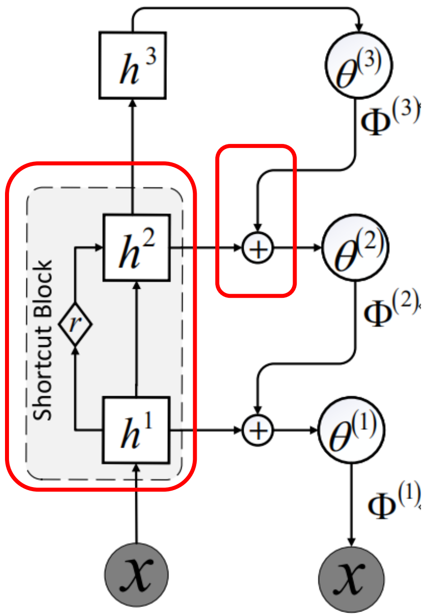
Our contribution

■ Encoder: Upward and Downward Encoder Networks

$$\boldsymbol{h}_j^{(l)} = \boldsymbol{h}_j^{(l-1)} + \mathrm{MLP}(\boldsymbol{h}_j^{(l-1)}) \qquad q(\boldsymbol{\theta}_j^{(l)} | \boldsymbol{\Phi}^{(l+1)}, \boldsymbol{h}_j^{(l)}, \boldsymbol{\theta}_j^{(l+1)}) = \mathrm{Weibull}(\boldsymbol{k}_j^{(l)}, \boldsymbol{\lambda}_j^{(l)})$$

$$\hat{\boldsymbol{k}}_j^{(l)} = \mathrm{Relu}(\mathrm{Linear}(\boldsymbol{h}_j^{(l)})) \qquad \boldsymbol{k}_j^{(l)} = \mathrm{Softplus}(\mathrm{Linear}(\boldsymbol{\Phi}^{(l+1)}\boldsymbol{\theta}_j^{(l+1)} \oplus \hat{\boldsymbol{k}}_j^{(l)}))$$

$$\hat{\boldsymbol{\lambda}}_j^{(l)} = \mathrm{Relu}(\mathrm{Linear}(\boldsymbol{h}_j^{(l)})) \qquad \boldsymbol{\lambda}_j^{(l)} = \mathrm{Softplus}(\mathrm{Linear}(\boldsymbol{\Phi}^{(l+1)}\boldsymbol{\theta}_j^{(l+1)} \oplus \hat{\boldsymbol{\lambda}}_j^{(l)}))$$



■ Weibull Reparameterization

$$\text{Weibull PDF: } P(x \,|\, k, \lambda) = \frac{k}{\lambda^k} x^{k-1} e^{-(x/\lambda)^k}$$

$$x = \lambda(-\ln(1-\varepsilon))^{1/k}, \quad \varepsilon \sim \mathrm{Uniform}(0,1).$$

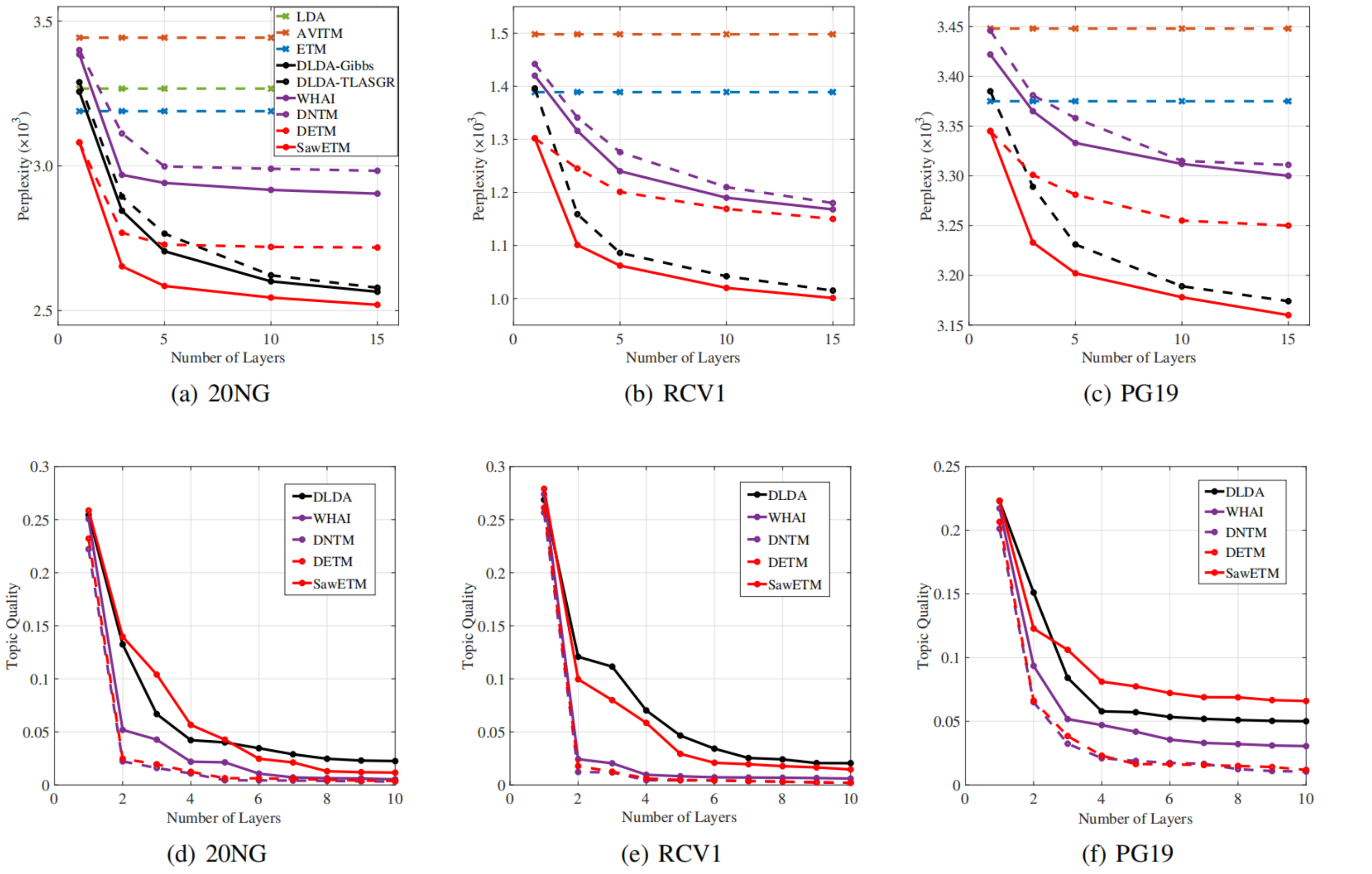
⇩ **Approximate**

$$\text{Gamma PDF: } P(x \,|\, \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}$$

- ✓ Fast in out-of-sample prediction
- ✓ Parallel scalable inference
- ✓ Alleviate the posterior collapse problem at the higher layer

Experiment

- Per-heldout-word Perplexity & Topic Quality



Perplexity

Topic Quality

Figure 2. (a)-(c): Comparison of per-heldout-word perplexity (the lower the better). (d)-(f): Comparison of topic quality (the higher the better).

Experiment

Document Clustering

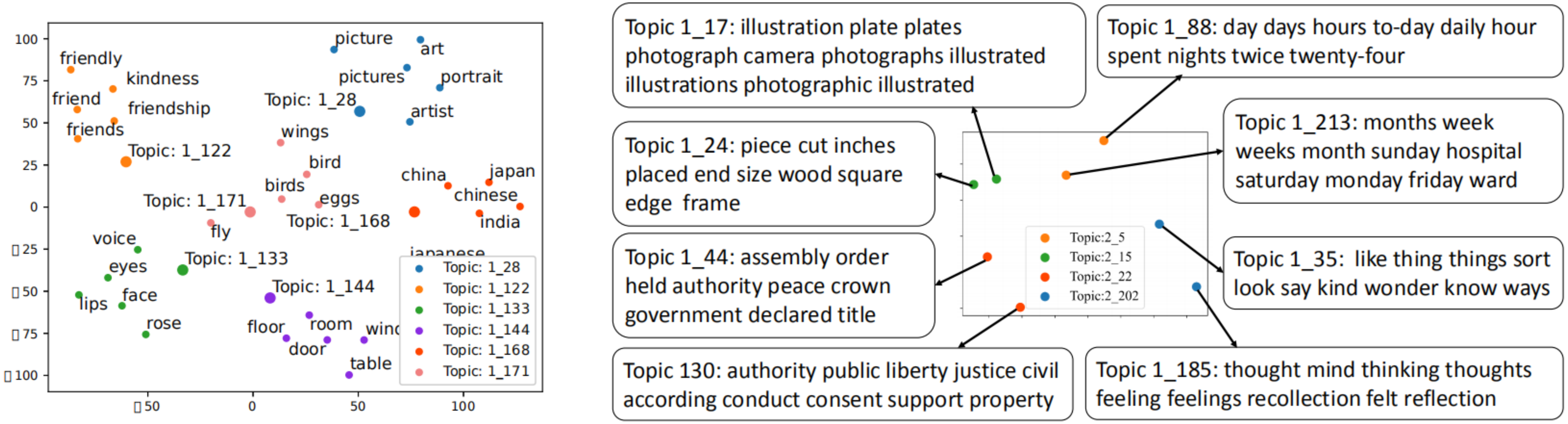
Table 1. Results of AC and NMI for document clustering task.

Model	Layer	20News		R8	
		AC	NMI	AC	NMI
LDA	1	46.52	45.15	51.41	40.47
AVITM	1	48.31	46.33	52.43	41.20
ETM	1	49.79	48.40	55.34	41.28
PGBN	1	46.62	45.43	51.67	40.76
PGBN	5	48.33	46.51	54.21	41.21
WHAI	1	49.43	46.56	57.86	42.31
WHAI	5	49.51	46.98	60.45	43.98
DNTM	1	49.17	46.32	57.58	42.12
DNTM	5	49.25	46.79	59.93	43.90
DETM	1	50.24	48.69	61.21	43.45
DETM	5	50.33	48.87	61.86	44.12
SawETM	5	**51.25**	**50.77**	**63.82**	**45.90**

Experiment

Qualitative Analysis: Visualization of Embedding Space

- T-SNE visualization of word embedding and topic embedding



(a) Word Embeddings (b) Topic Embedding

Figure 3. t-SNE visualisation of (a) word embeddings, which we choose the top ten words for each topic at layer one and (b) topic embeddings, which we choose the top two sub topics for each topic at layer two. (Note that the Topic: t_j denotes the j^{th} topic at t^{th} layer.)

Experiment

■ Qualitative Analysis : Hierarchical Structure of Topic Model

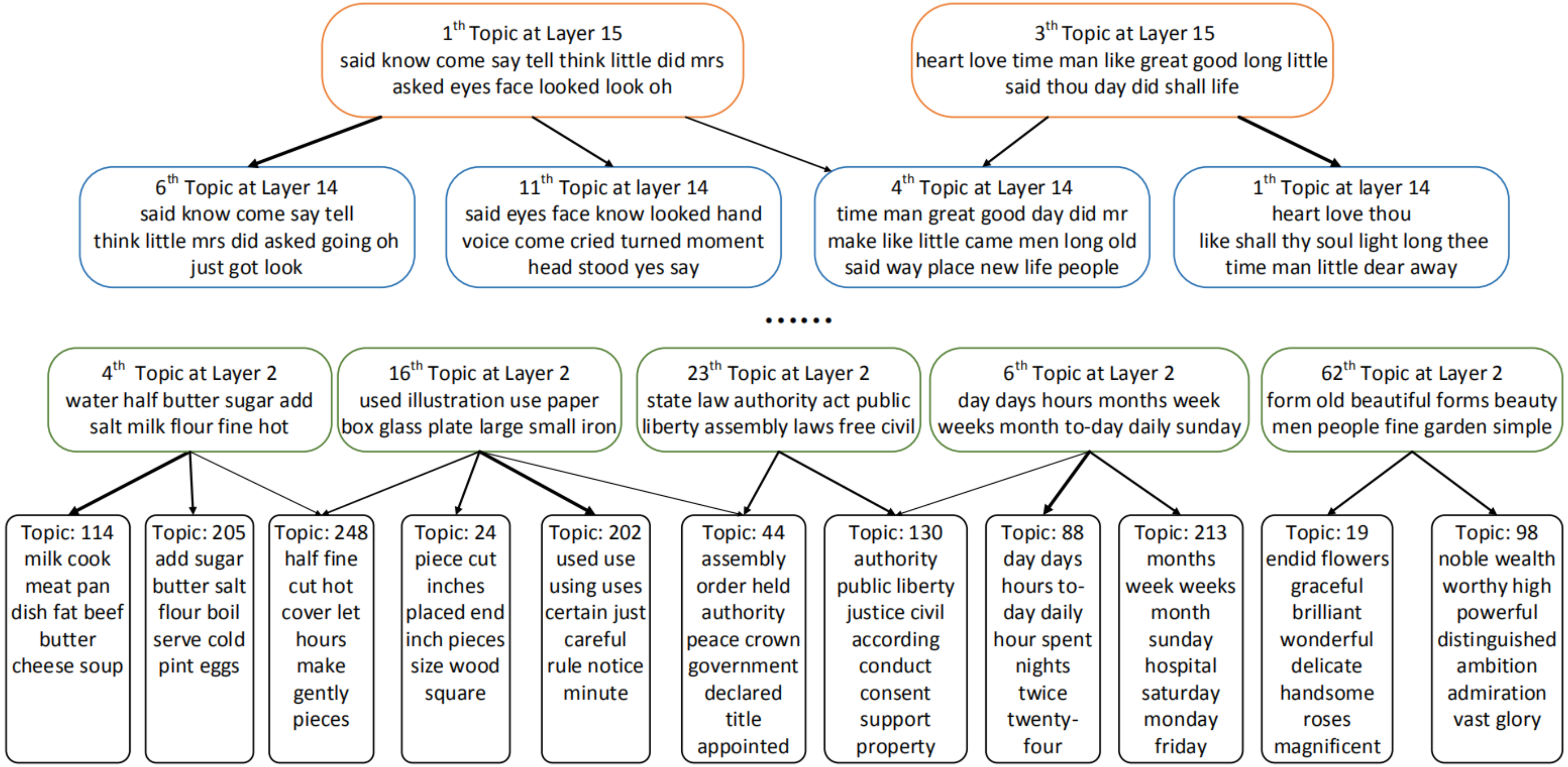


Figure 4. An example of hierarchical topics learned from PG-19 by a 15-layer SawETM, We only show example topics at the top two layers and bottom two layers.

西安電子科技大學
XIDIAN UNIVERSITY

THE UNIVERSITY OF TEXAS AT AUSTIN
McCOMBS SCHOOL OF BUSINESS

THANK YOU!

Email address: xd_zhibin@163.com

Code available: https://github.com/BoChenGroup/SawETM