Integrating Semantics and Neighborhood Information with Graph-Driven Generative Models for Document Retrieval

Zijing Ou, Qinliang Su, Jianxing Yu, Bang Liu, Jingwen Wang, Ruihui Zhao,
Changyou Chen and Yefeng Zheng

Sun Yat-sen University, University Montreal, Tencent Jarvis Lab
and SUNY at Buffalo

ouzj@mail2.sysu.edu.cn

Representation learning on text

- Generative framework

$$p_\theta(x, z) = p_\theta(x|z)p(z)$$

$$x = \{w_1, w_2, \dots, w_{|x|}\}$$

One-hot representation

z is d dimensional vector representation

- Softmax decoder

$$p_\theta(w_i|z) = \frac{\exp(z^T E w_i + b_i)}{\sum_{j=1}^{|V|} \exp(z^T E w_j + b_j)}$$

$p_\theta(x|z) = \prod_{i=1}^{|x|} p_\theta(w_i|z)$ (iid. assumption)

Variational inference

- Evidence lower bound

$$\log p(x) = ELBO + KL(q_\phi(z|x)||p(z))$$

$$\geq ELBO = E_{q_\phi(z|x)}\left[\log \frac{p_\theta(x|z)p(z)}{q_\phi(z|x)}\right]$$

- Consistent learning & inference

$$\mathcal{L} = E_{p(x)}E_{q_\phi(z|x)}\left[\log \frac{p_\theta(x|z)p(z)}{q_\phi(z|x)}\right]$$

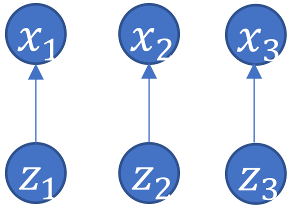
$$q_\phi(z|x) \coloneqq \mathcal{N}\left(z \middle| \mu_\phi(x), diag\left(\sigma_\phi^2(x)\right)\right)$$

Mean field inference

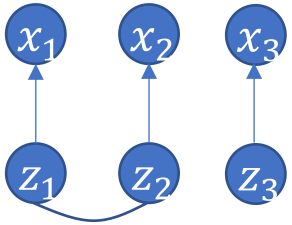
→ Independent among data

Our focus: breaking the independence assumption!!!

Data dependence prior



x_1 x_2 x_3



$$p(X,Z) = p_\theta(X|Z)p_I(Z)$$

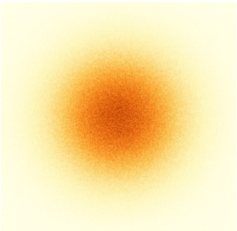
$$X = \{x_1, x_2, x_3\}$$

$$p(X,Z) = p_\theta(X|Z)p_G(Z)$$

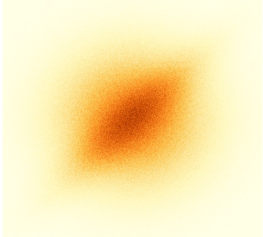
data independence prior

$$p_I(Z) = \mathcal{N}(Z; I \otimes I_d)$$

denote as Σ_I



$p_I(Z)$



$p_G(Z)$

data dependence prior

$$p_G(Z) = \mathcal{N}(Z;\ (I + \lambda A) \otimes I_d)$$

denote as Σ_G

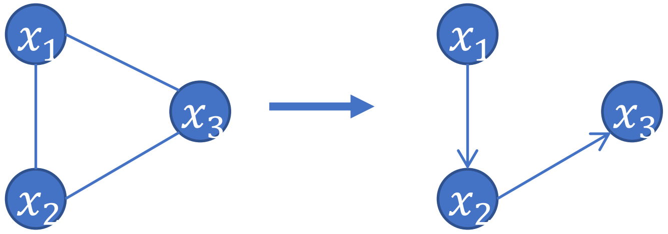
$$\Sigma_I = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

$$\Sigma_G = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

$$\mathcal{L} \coloneqq E_{q_\phi(Z|X)}[\log p_\theta(X|Z)] - KL(q_\phi(Z|X)||p_G(Z))$$

"curse-of-dimensionality" $O((Nd)^3)$

Spanning-tree approximations



$$p_{\mathrm{T}}(Z) = \prod_{i\in\mathcal{V}} p_G(z_i) \prod_{(i,j)\in\mathcal{E}} \frac{p_G(z_i,z_j)}{p_G(z_i)p_G(z_j)}$$

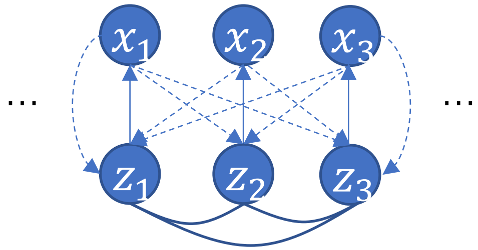
$$q_{\mathrm{T}}(Z|X) = \prod_{i\in\mathcal{V}} q_\phi(z_i|x_i) \prod_{(i,j)\in\mathcal{E}} \frac{q_\phi(z_i,z_j|x_i,x_j)}{q_\phi(z_i|x_i)q_G(z_j|x_j)}$$

Covariance of $\boldsymbol{q_\phi(z_i, z_j|x_i, x_j)}$:

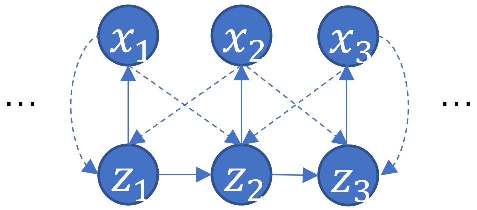
$\gamma_{ij} \in (0, 1)$: positive correlated

$$\begin{bmatrix} diag(\sigma_i^2) & diag(\gamma_{ij}\odot\sigma_i\odot\sigma_j) \\ diag(\gamma_{ij}\odot\sigma_i\odot\sigma_j) & diag(\sigma_j^2) \end{bmatrix}$$

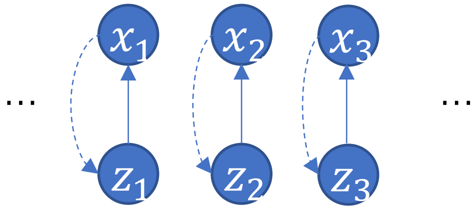
Spanning-tree approximations



Graph driven



 Tree approximation



Mean field

$$\mathcal{L}_T = \sum_{i \in \mathcal{V}} E_{q_\phi}[\log p_\theta(x_i|z_i)] - KL(q_\phi(z_i)||p_G(z_i))$$

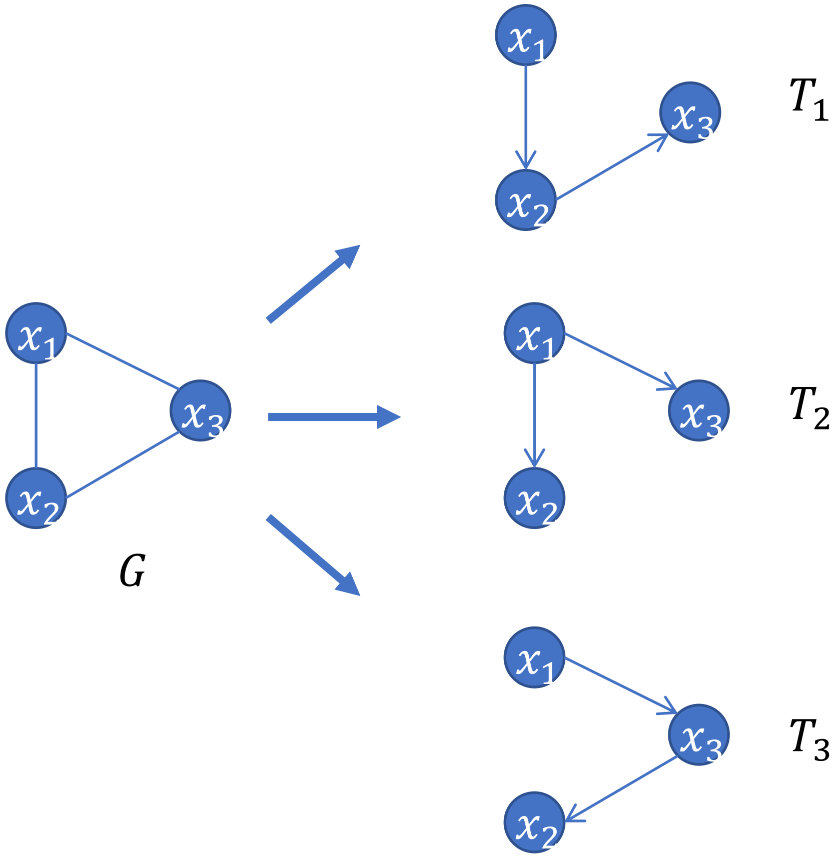
$$= \sum_{(i,j) \in \mathcal{E}} \left(KL\left(q_\phi(z_i, z_j)||p_G(z_i, z_j)\right) - KL\left(q_\phi(z_i)||p_G(z_i)\right) - KL\left(q_\phi(z_j)||p_G(z_j)\right) \right)$$

Extending to multiple trees

$$p_{\mathrm{MT}}(Z) = \frac{1}{M} \sum_{T \in T_G} p_T(Z)$$

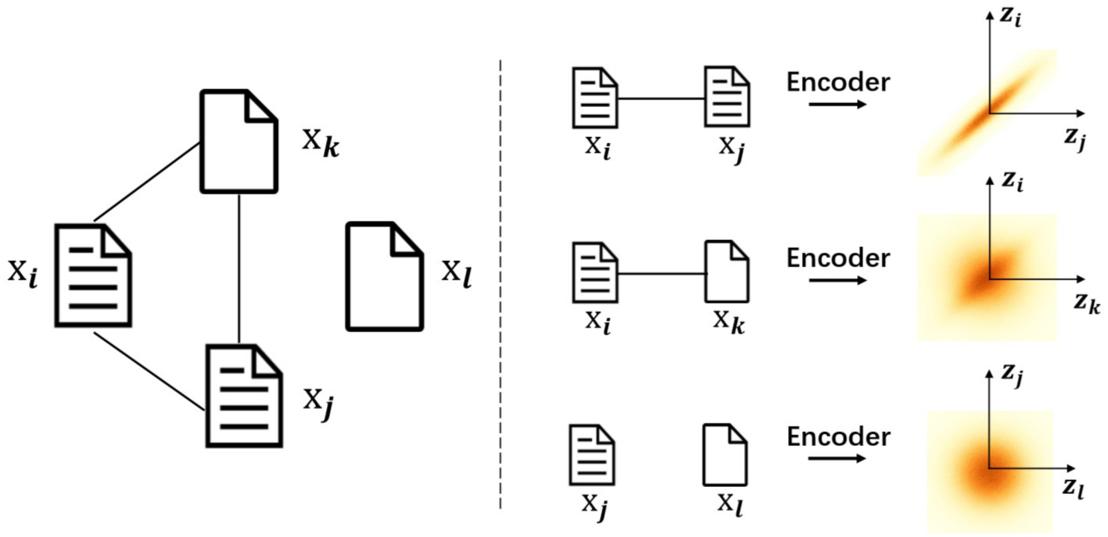
$$q_{\mathrm{MT}}(Z) = \frac{1}{M} \sum_{T \in T_G} q_T(Z|X)$$

$$\widetilde{\mathcal{L}_{MT}} = \frac{1}{M} \mathcal{L}_T$$



$$T_G = \{T_1, T_2, T_3\}$$

Details of modeling



- **Variational Encoder** $q_\varphi(z_i|x_i)$
  - take single document as input, and outputs the mean and variance of Gaussian distribution $[\mu_i; \sigma_i^2] = f_\varphi(x_i)$.
- **Correlation Encoder**
  - take pairwise documents as input, and outputs the correlation coefficient $\gamma_{ij} = f_\varphi(x_i, x_j)$.
- **Generative Decoder** $p_\theta(x_i|z_i)$
  - take the latent variable z_i as input and output the document x_i.

Documents retrieval results

Method	Reuters				TMC				20Newsgroups				Avg
	16bits	32bits	64bits	128bits	16bits	32bits	64bits	128bits	16bits	32bits	64bits	128bits	
SpH	0.6340	0.6513	0.6290	0.6045	0.6055	0.6281	0.6143	0.5891	0.3200	0.3709	0.3196	0.2716	0.5198
STH	0.7351	0.7554	0.7350	0.6986	0.3947	0.4105	0.4181	0.4123	0.5237	0.5860	0.5806	0.5443	0.5662
VDSH	0.7165	0.7753	0.7456	0.7318	0.6853	0.7108	0.4410	0.5847	0.3904	0.4327	0.1731	0.0522	0.5366
NbrReg	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	0.4120	0.4644	0.4768	0.4893	0.4249
NASH	0.7624	0.7993	0.7812	0.7559	0.6573	0.6921	0.6548	0.5998	0.5108	0.5671	0.5071	0.4664	0.6462
GMSH	0.7672	0.8183	0.8212	0.7846	0.6736	0.7024	0.7086	0.7237	0.4855	0.5381	0.5869	0.5583	0.6807
AMMI	0.8173	0.8446	0.8506	0.8602	0.7096	0.7416	0.7522	0.7627	0.5518	0.5956	0.6398	0.6618	0.7323
CorrSH	0.8212	0.8420	0.8465	0.8482	0.7243	0.7534	0.7606	0.7632	**0.5839**	0.6183	0.6279	0.6359	0.7355
SNUH	**0.8320**	**0.8466**	**0.8560**	**0.8624**	**0.7251**	**0.7543**	**0.7658**	**0.7726**	0.5775	**0.6387**	**0.6646**	**0.6731**	**0.7474**

- Perform on three datasets consistently better than baselines in most experimental settings.

Effects of correlated prior&posterior

Ablation Study		16bits	32bits	64bits	128bits
Reuters	$SNUH_{ind}$	0.7823	0.8094	0.8180	0.8385
	$SNUH_{prior}$	0.8043	0.8295	0.8431	0.8460
	SNUH	**0.8320**	**0.8466**	**0.8560**	**0.8624**
TMC	$SNUH_{ind}$	0.6978	0.7307	0.7421	0.7526
	$SNUH_{prior}$	0.7177	0.7408	0.7518	0.7528
	SNUH	**0.7251**	**0.7543**	**0.7658**	**0.7726**
NG20	$SNUH_{ind}$	0.4806	0.5503	0.6017	0.6060
	$SNUH_{prior}$	0.5443	0.6071	0.6212	0.6014
	SNUH	**0.5775**	**0.6387**	**0.6646**	**0.6731**

$\mathrm{SNUH}_{\mathrm{ind}}$

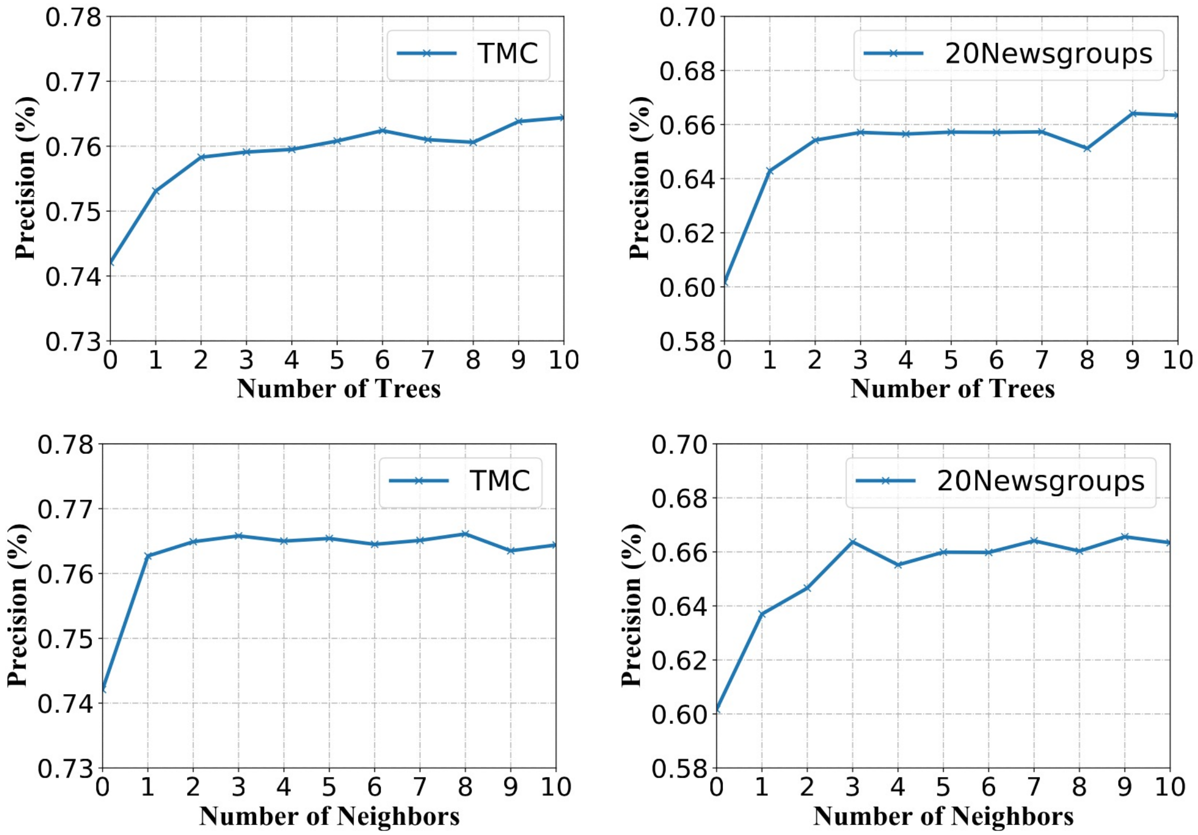
without considering correlation (independent)

$SNUH_{prior}$

only consider correlation in the prior

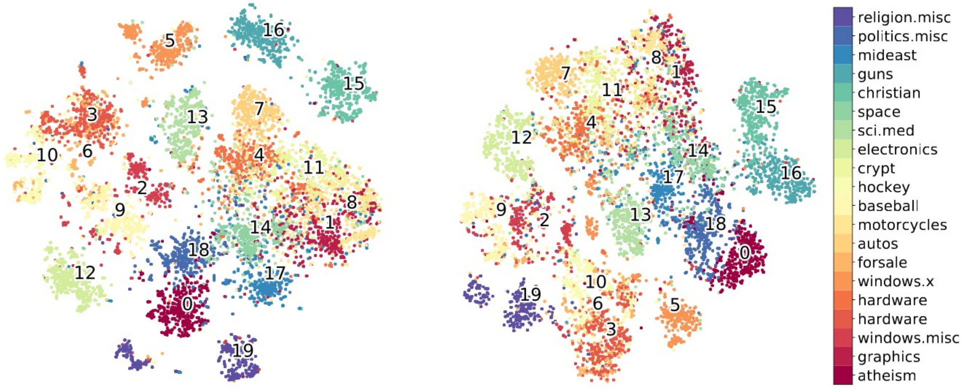
- By taking the **correlations** into account in the prior and posterior, significant **improvements** of SNUH can be observed.

Impact of the number of trees



- Compared to not using any correlation, **one tree** alone can bring significant performance gains.

Case study



(a) SNUH (b) AMMI

Distance	Category	Title/Subject
query	**hockey**	**NHL PLAYOFF RESULTS FOR GAMES PLAYED 4-21-93**
1	hockey	NHL PLAYOFF RESULTS FOR GAMES PLAYED 4-19-93
10	hockey	NHL Summary parse results for games played Thur, April 15, 1993
20	hockey	AHL playoff results (4/15)
50	forsale	RE: == MOVING SALE ===
70	hardware	Re: Quadra SCSI Problems?
90	politics.misc	Re: Employment (was Re: Why not concentrate on child molesters?

Looking forward

Structured representation learning

- Expressive prior & posterior
- Combine with advances in GNNs
- Graphical models meet Deep Learning (GNN as message passing)

Thank you!