Supervised Hierarchical Dirichlet Processes with Variational Inference

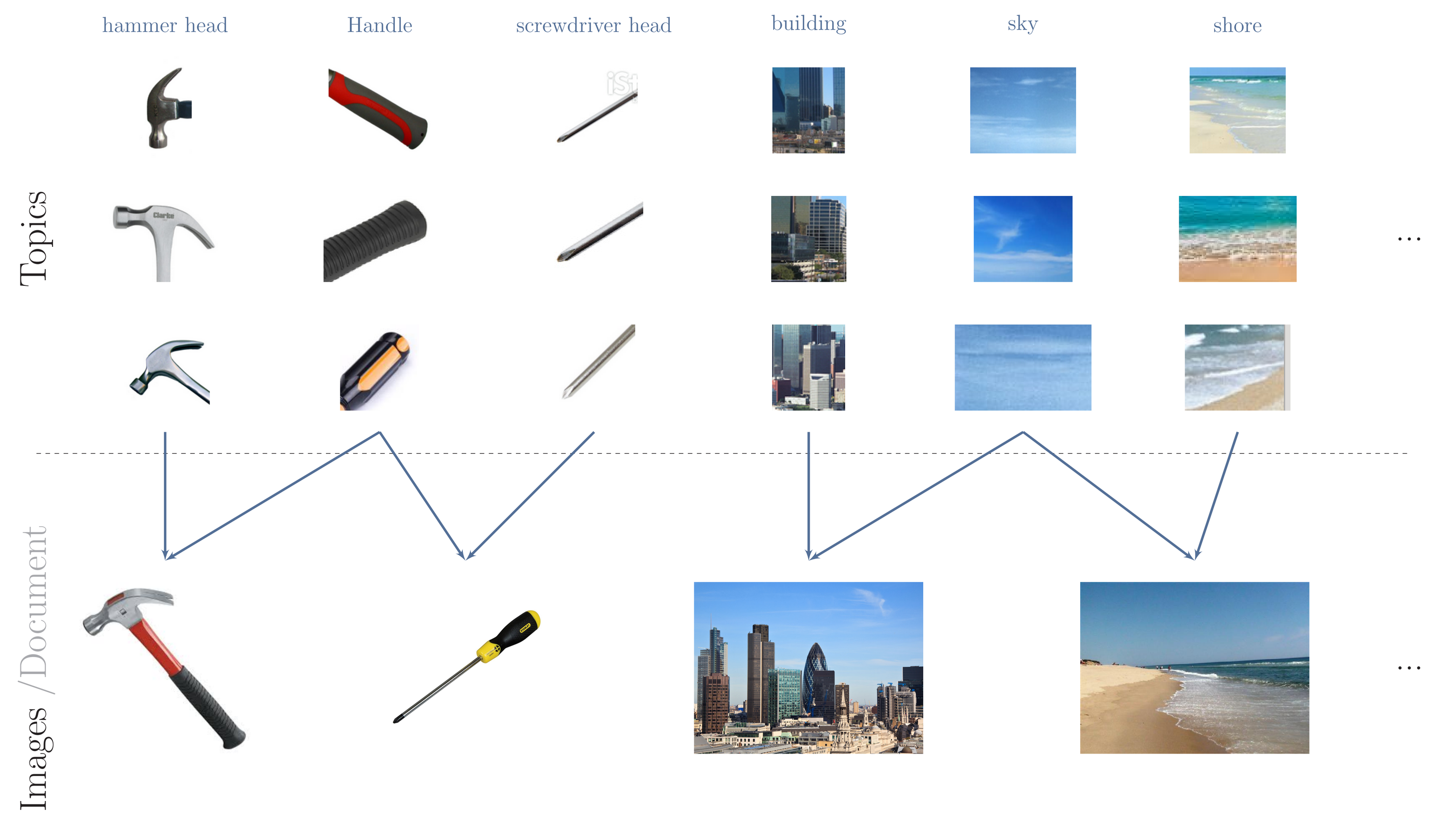
Cheng Zhang, Carl Henrik Ek, Xavi Gratal, Florian T. Pokorny, Hedvig Kjellström

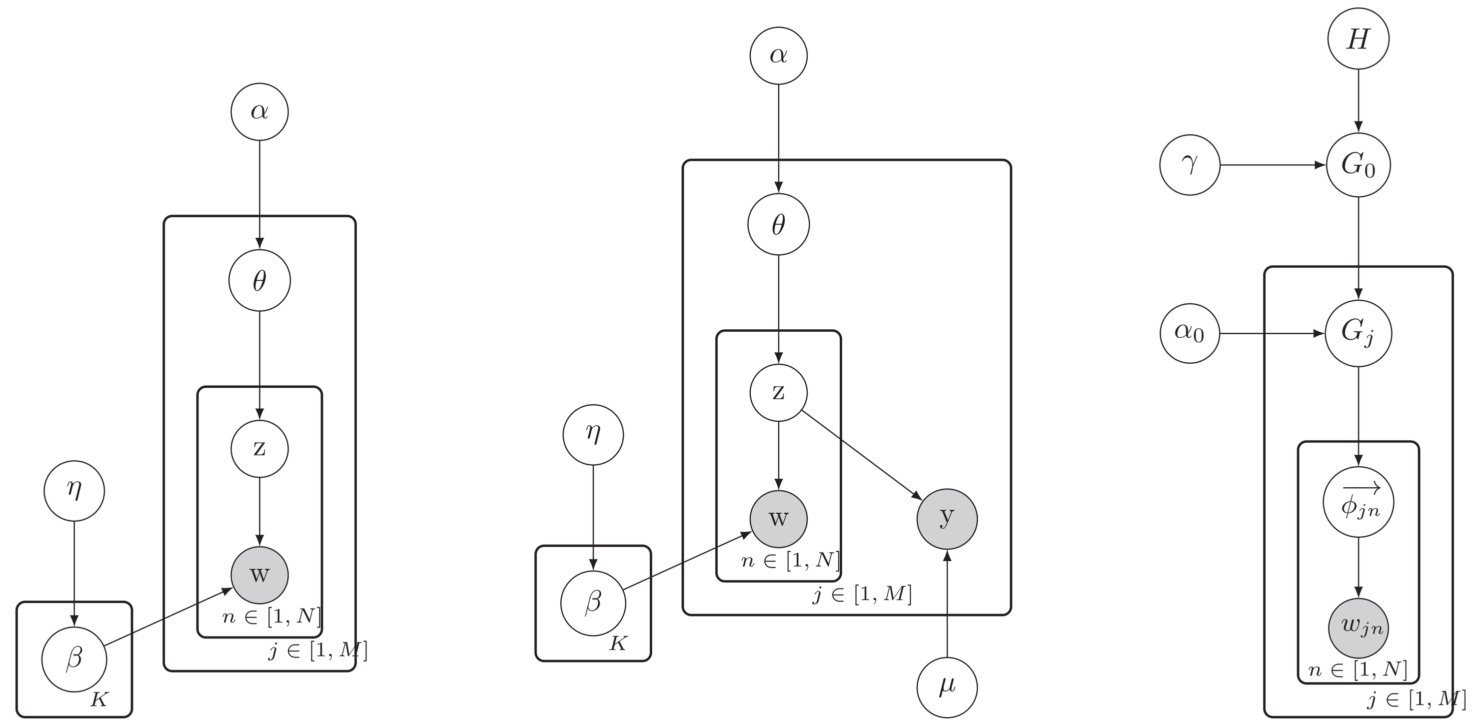
{chengz, chek, javiergm, fpokorny, hedvig}@kth.se

Abstract

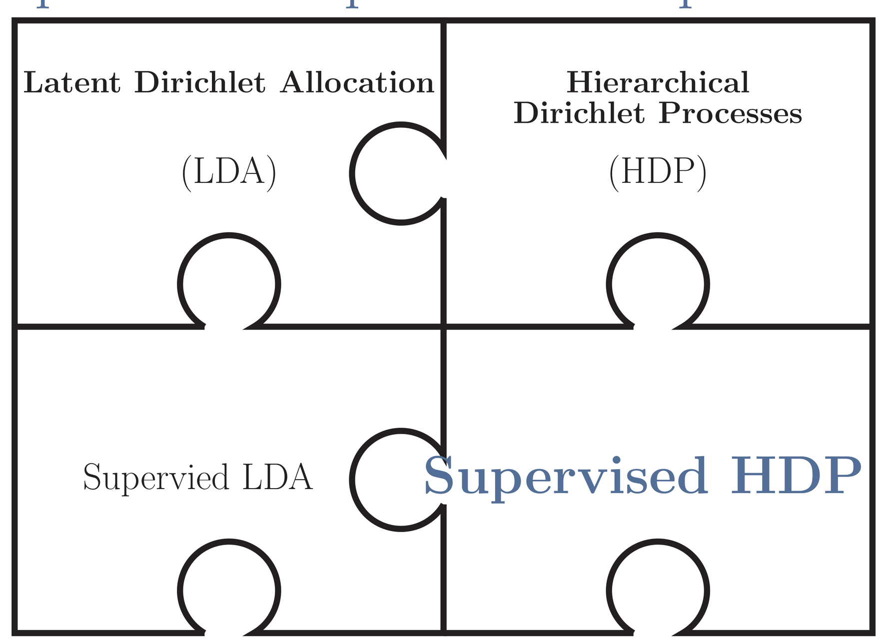
We present an extension to the Hierarchical Dirichlet Process (HDP), which allows for the inclusion of supervision. Our model marries the non-parametric benefits of HDP with those of Supervised Latent Dirichlet Allocation (SLDA) to enable learning the topic space directly from data while simultaneously including the labels within the model. The proposed model is learned using variational inference which allows for the efficient use of a large training dataset. We also present the online version of variational inference, which makes the method scalable to very large datasets. We show results comparing our model to a traditional supervised parametric topic model, SLDA, and show that it outperforms SLDA on a number of benchmark datasets.

Motivation



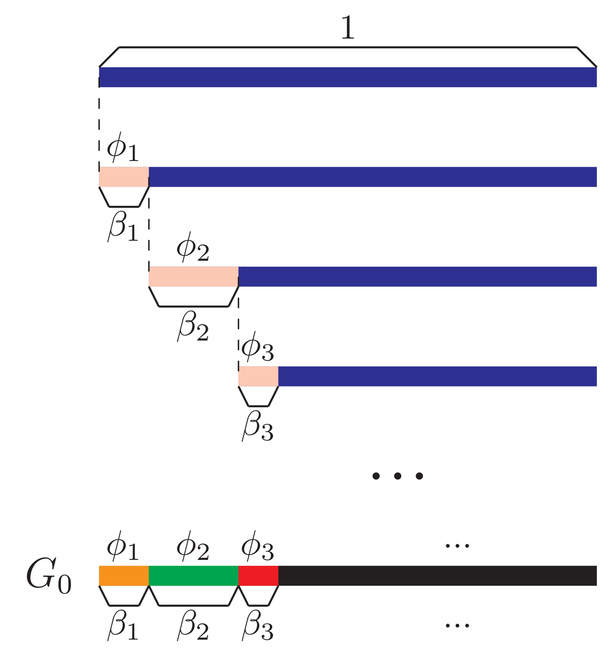


Supervised Nonparametric Topic Model



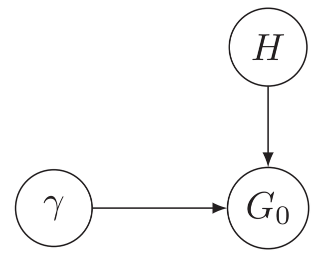
Model

DP

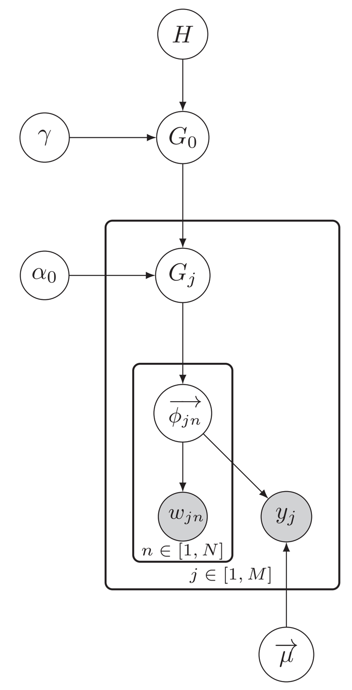


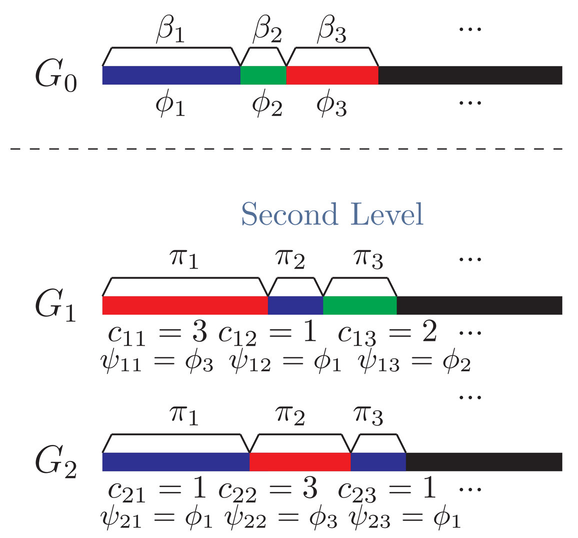
$G_0 \sim DP(H, \gamma)$

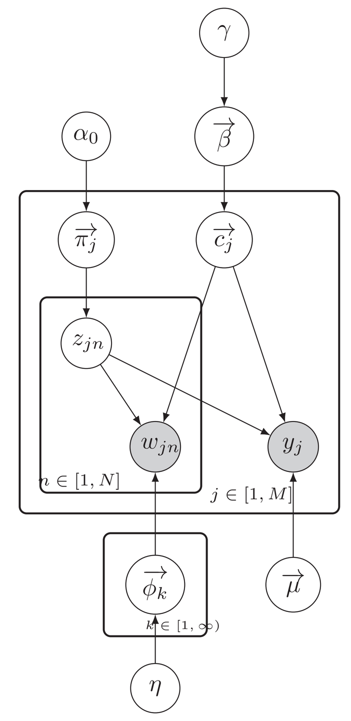
$\beta'_k \sim Beta(1, \gamma)$,
$\beta_k = \beta'_k \prod_{l=1}^{k-1}(1-\beta'_l)$,
$\phi_k \sim H$,
$G_0 = \sum_{k=1}^{\infty} \beta_k \delta_{\phi_k}$.
β:s are general decreasing but not strictly



SHDP







$$c_{jt} \sim Mult(\beta)$$

Softmax Function

$$p(y_j|\bar{\boldsymbol{\theta}}_j, \boldsymbol{\mu}) = \frac{\exp(\boldsymbol{\mu}_{y_j}^T \bar{\boldsymbol{\theta}}_j)}{\sum_{l=1}^{C} \exp(\boldsymbol{\mu}_l^T \bar{\boldsymbol{\theta}}_j)}, \text{ where } \bar{\boldsymbol{\theta}}_j = \frac{1}{N}\sum_{n=1}^{N} \Theta_{jn} = \frac{1}{N}\sum_{n=1}^{N} \mathfrak{c}_{j z_{jn}}$$

Classification: $\hat{y}_j = argmax_{y_j \in \{1,\dots,C\}} \mathbb{E}[\boldsymbol{\mu}_{y_j}^T \bar{\boldsymbol{\theta}}_j]$

Variational Inference

$$q(\boldsymbol{\beta}', \boldsymbol{\pi}', \mathbf{c}, \mathbf{z}, \boldsymbol{\phi}) = q(\boldsymbol{\beta}')q(\boldsymbol{\pi}')q(\mathbf{c})q(\mathbf{z})q(\boldsymbol{\phi}),$$

The supervision term in the lower bound:

$$\mathbb{E}_q[\log p(y_j|\mathbf{z}, \mathbf{c}, \boldsymbol{\mu})] \geq \boldsymbol{\mu}_{y_j}^T \left(\frac{1}{N}\sum_{n=1}^{N}\sum_{t=1}^{T} \boldsymbol{\rho}_{jt}\zeta_{jnt}\right)$$

$$-\log\left(\sum_{l=1}^{C}\prod_{n=1}^{N}\left(\sum_{e=1}^{K}\sum_{i=1}^{T} \rho_{jie}\zeta_{jni}\exp\left(\frac{1}{N}\mu_{le}\right)\right)\right).$$

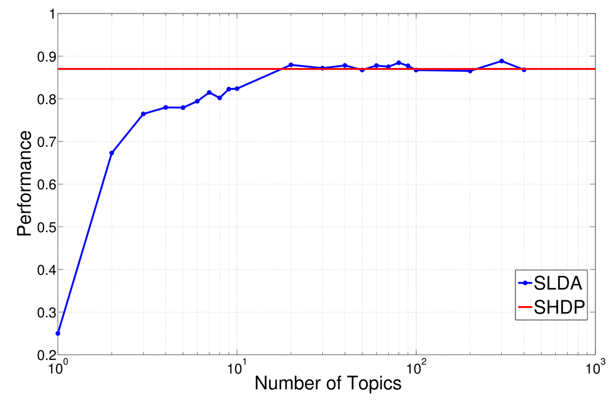
Online: $\mathcal{L} = \sum_j^M \mathcal{L}_j = \mathbb{E}_j[M\mathcal{L}_j]$.

Experiments

Scene Classification

	coast	forest	mountain	opencountry
coast	0.78	0	0	0.22
forest	0	0.95	0.03	0.02
mountain	0	0.01	0.89	0.09
opencountry	0.1	0.01	0.05	0.84

Confusion Matrix 86.68%

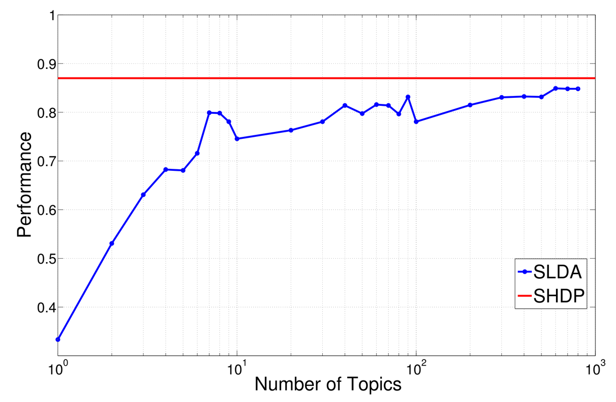


Scene Classification performance

Action Classification

	boxing	clapping	waving
boxing	0.85	0.1	0.05
clapping	0	1	0
waving	0.1	0.15	0.75

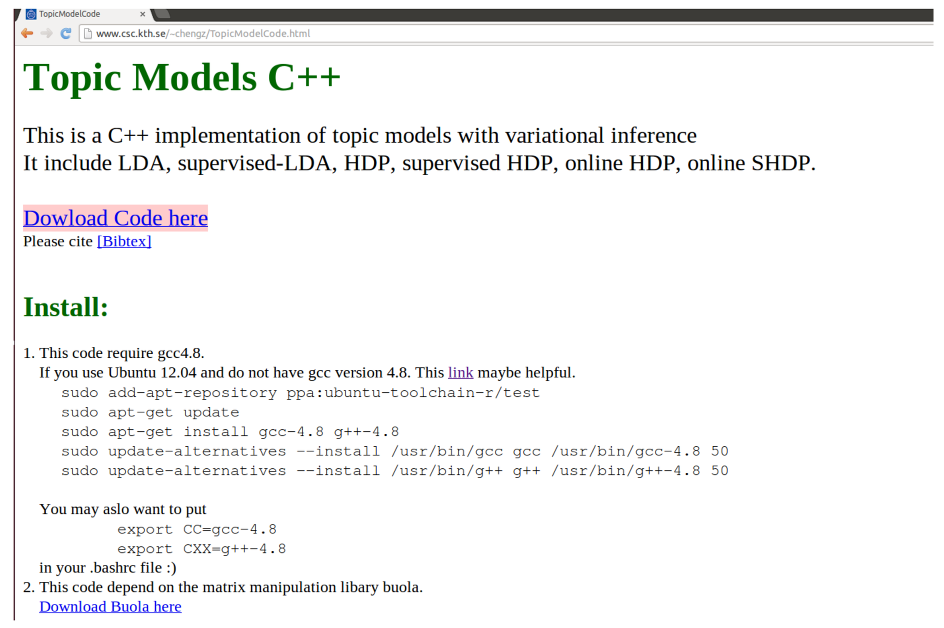
Confusion Matrix 86.67%



Action Classification performance

Code Available

http://www.csc.kth.se/~chengz/TopicModelCode.html



Acknowledgement

This research has been supported by the EU through TOMSY, IST-FP7- Collaborative Project-270436, and the Swedish Research Council (VR).

References

[1] D. M. Blei, A. Y. Ng and M. I. Jordan. *Latent Dirichlet Allocation* Journal of Machine Learning Research, 3:993-1022, 2003

[2] Y. W. Teh, M. I. Jordan, M. J. Beal and D.M. Blei. *Hierarchical Dirichlet Process* Journal of the American Statistical Association,101(476):1566-1581, 2006.

[3] C. Wang, J. Paisley and D. Blei. *Online Variational inference for the Hierarchical Dirichlet Process* In AISTATS, 2011

[4] D. M. Blei and J. D. McAuliffe. *Supervised topic models* In NIPS, 2007