Perspective Hierarchical Dirichlet Process for User-Tagged Image Modeling

Xin Chen[1], Xiaohua Hu[1], Yuan An[1], Zunyan Xiong[1], Tingting He[2], E.K. Park[3]

[1]College of Information Science and Technology, Drexel University, Philadelphia, PA 19104, USA
[2]Dept. of Computer Science at Central China Normal University, Wuhan, China
[3]California State University - Chico, Chico, CA 95929, USA

Outlines

- Introduction & Research Questions
- Background & Related Works
  - Framework of Image Feature Representation
  - Generative Models for Image Features and Text
- Developed Model and Evaluation
  - Perspective Hierarchical Dirichlet Process (pHDP)
  - Evaluations
- Conclusions

Flickr image tags as examples of social annotations

San Giorgo Maggiore

Would you like to comment?

Sign up for a free account, or sign in (if you're already a member).

Uploaded on April 19, 2009
by Suki.

Suki.'s photostream

Vistas de Europa (Set)

You are at the last photo.

82 items

browse

Tags

- Italia
- Italy
- Isla
- Isle
- paisaje
- Landscape
- calle
- street
- Europa
- europe
- nikon
- D40X
- Venecia
- Venice
- agua
- water
- mar
- sea

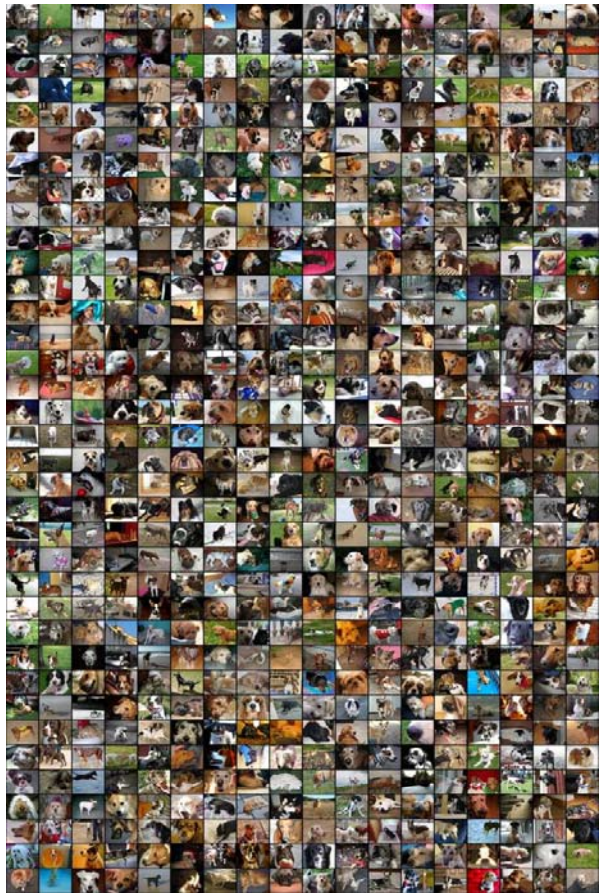
Illustration of Flickr image tags and the mapping to different social tagging classification schemas

ID: 08715
Title: So Far Away (_DSC9012)
Tags:
Lake, plant life, water, sky, 2007, Malaysia, Asia, Nikon, d50, landscape, 200mm, impressed beauty, vivid, an awesome shot vacation, holiday, travel, trip diamond class photographer, excellent photographer awards

Sen et al.[7]	Bischoff et al.[6]	Examples
Factual	Topic	Lake, plant life, water, sky
	Time	2007
	Location	Malaysia, Asia
	Type	Nikon, d50, landscape, 200mm
	Author/Owner	N/A
Subjective	Opinions/Qualities	impressed beauty, vivid, an awesome shot
Personal	Usage context	vacation, travel
	Self reference	diamond class photographer, excellent photographer awards

Objective: build models for the user-tagged image and achieve automatic image tagging

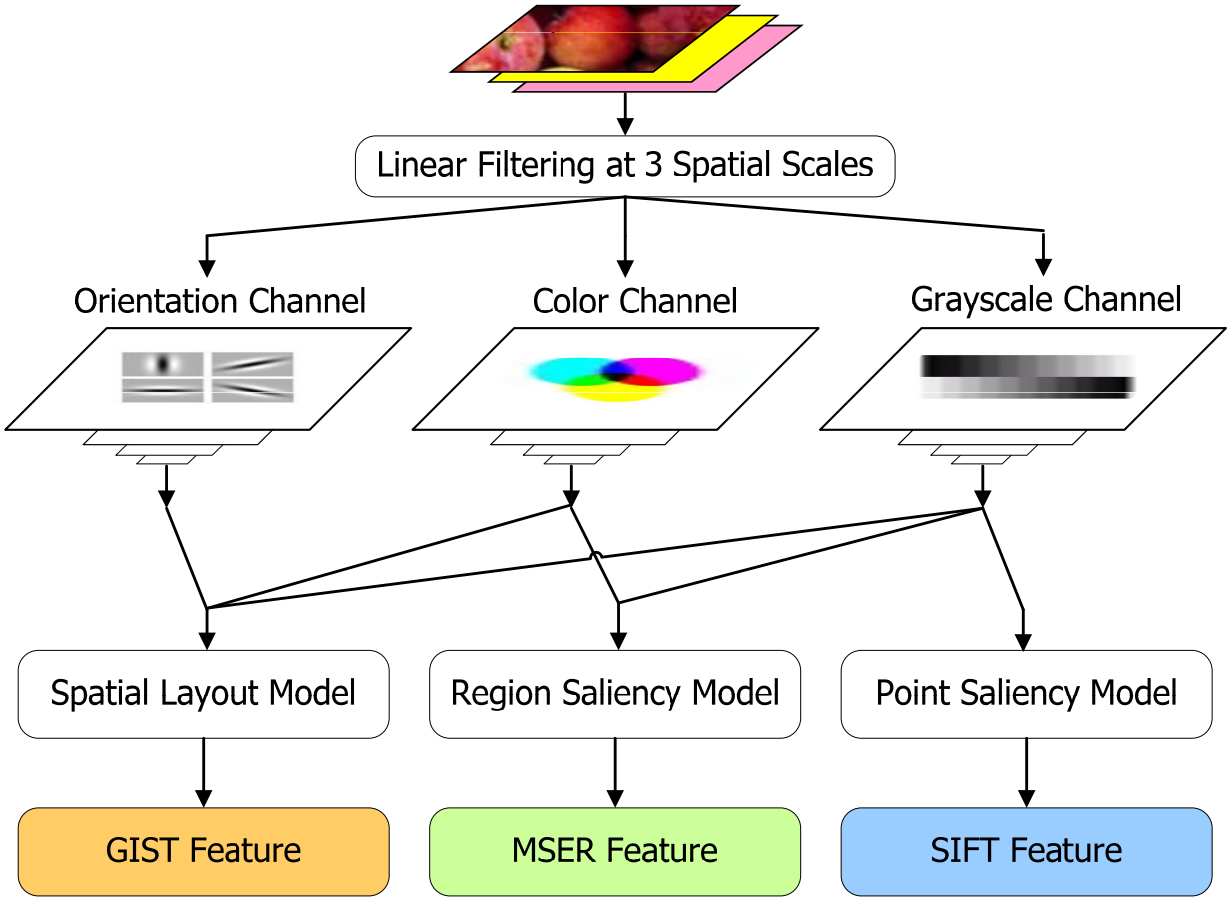
- Manual image tagging is time-consuming, laborious and expensive.
- User-tagged images not only provide insight on correlation between image content and tags, but also provide valuable contextual information of users' tagging preference which can be utilized to customize automatic image tagging for different users.
- Breakthroughs in automatic image tagging will help to organize the massive amount of digital images, promote developing and studying of image storage and retrieval systems, and serve for other applications such as online image-sharing.



Outlines

- Introduction & Research Questions
- Background & Related Works
  - Framework of Image Feature Representation
  - Generative Models for Image Features and Text
- Developed Model and Evaluation
  - Perspective Hierarchical Dirichlet Process (pHDP)
  - Evaluations
- Conclusions

Proposed image representation framework

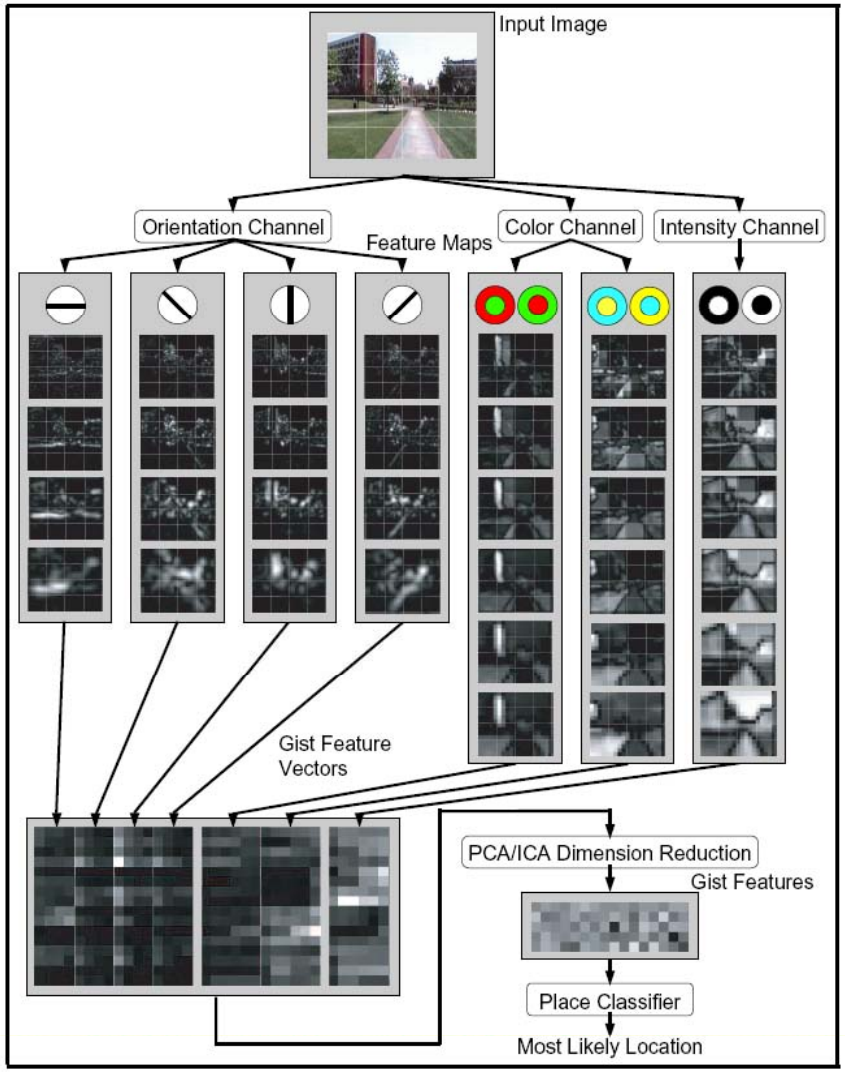


Outlines

- Introduction & Research Questions
- Background & Related Works
  - Framework of Image Feature Representation
  - Generative Models for Image Features and Text
- Developed Model and Evaluation
  - Perspective Hierarchical Dirichlet Process (pHDP)
  - Evaluations
- Conclusions

Holistic Image Representation - GIST Features (Siagian and Itti, 2007)

- The holistic image representation derived from the low resolution spatial layout not only provides a coarse context of image but also provides compact summarization of image's statistics and semantics.
- In practice, we extract the GIST features as a compact representation of image scene.
- The total number of raw GIST features per image is 714 (34 feature maps time 21 grids in a total of 3 scales). We reduce the dimension using principal component analysis (PCA) to a more practical number 100 (still preserving most image variance).



(Siagian and Itti, 2007)

Region Saliency Model - Maximally Stable Extremal Regions (MSER) Features (Matas et al. 2002)

- MSERs is a highly efficient region detector. The idea origins from thresholdings in image color/intensity space *I*. The thresholding yields a binary image E_t as follows:

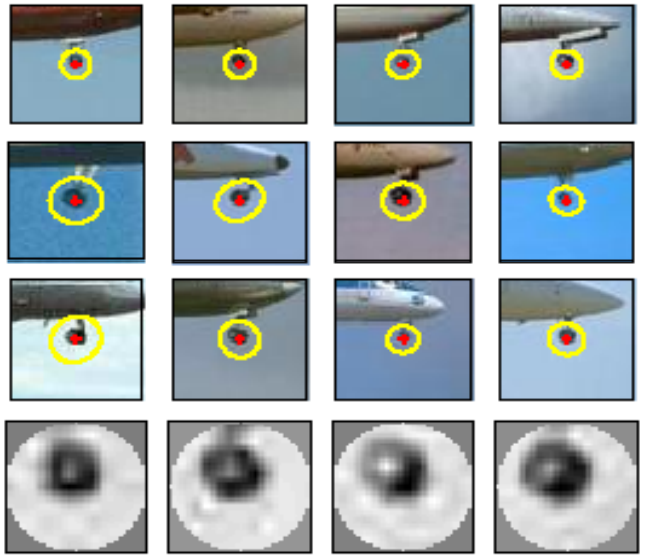
$$E_t(\mathbf{x}) = \begin{cases} 1 & \text{if } I(\mathbf{x}) \geq t \\ 0 & \text{otherwise.} \end{cases}$$

- An extremal region is maximally stable when the area (or the boundary length) of the segment changes the least with respect to the threshold.

The set of MSERs is closed under continuous geometric transformations and is invariant to affine intensity changes.

Quantifying the image parts in a continuous space

- Image patches containing salient parts are rotated to canonical angle and adjust to uniform size (known as normalized patches).
- Principal component analysis (PCA) is performed on normalized patches to obtain feature representation

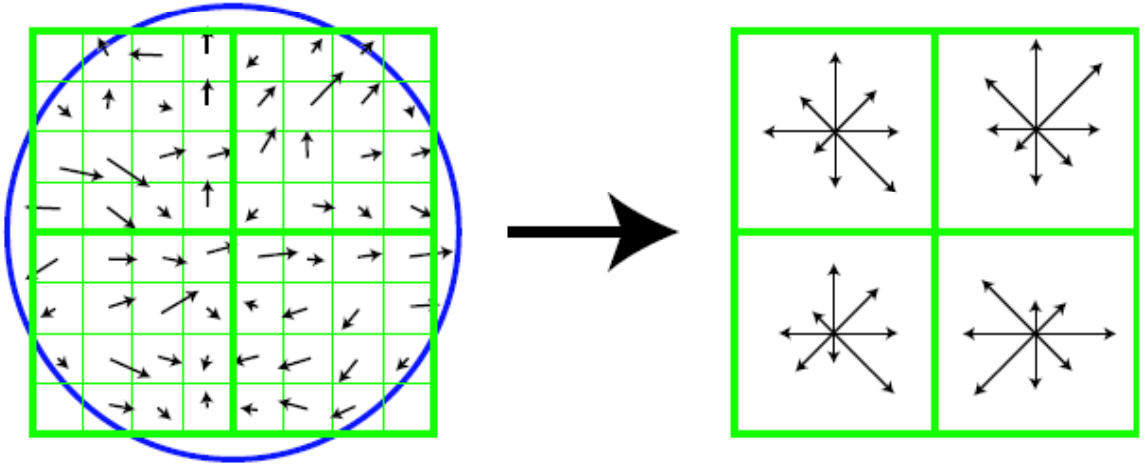


Adjusting image patches to uniform size

Finally, the appearance of each patch (which is n × n matrix) is quantified as a feature vector of the first k (typically 20-50) principal components

Point Saliency Model - Scale Invariant Feature Transform (SIFT) Features (Lowe, 2004)

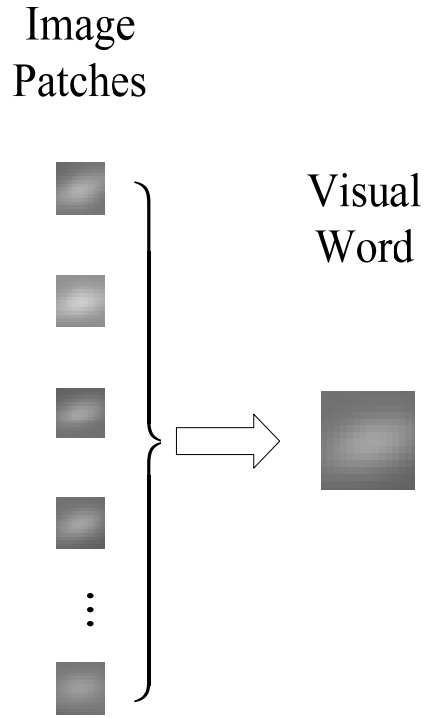
- Image patches containing salient points are rotated to a canonical orientation and divided into cells. Each cell is represented as an 8-dimension feature vector according to the gradient magnitude in eight orientations.
- Compared to other descriptors, the SIFT descriptor is more robust and invariant to rotation and scale/luminance changes.



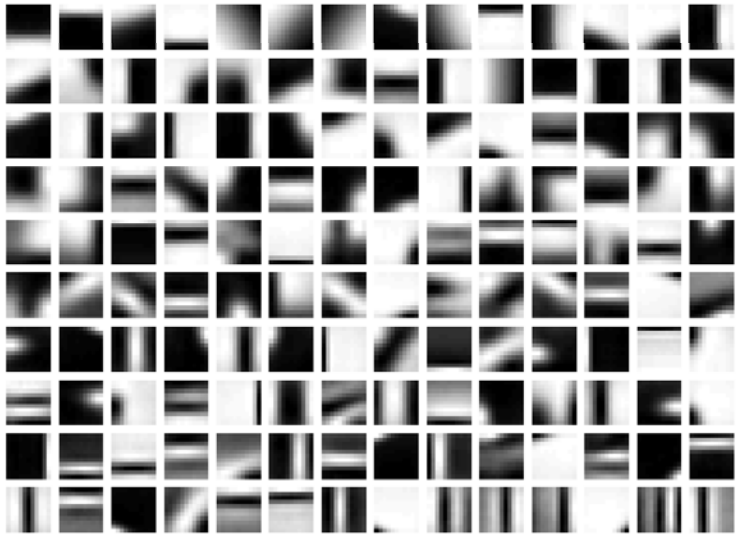
The SIFT descriptor of salient points (2×2 cells) (Lowe, 2004)

Grouping similar local descriptors into visual words

- Typically, the K-Mean clustering algorithm is used to cluster the descriptors of extracted image patches into visual words and establish a code book of visual words for a specific image collection.



Each key-point assigned the closest cluster center

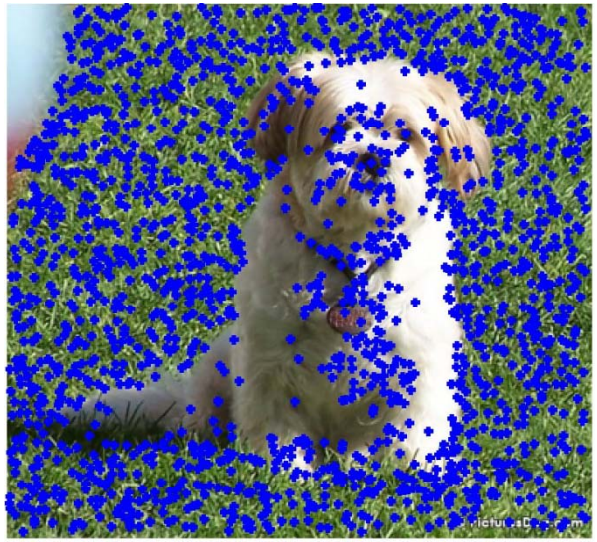


Code book of visual words (Sivic, 2003) and (Fei-Fei et al. 2005)

Summary: image represented by salient points and regions

- Represent image by SIFT descriptors and MSER features

SIFT
(key-points)



MSER
(parts)

Outlines

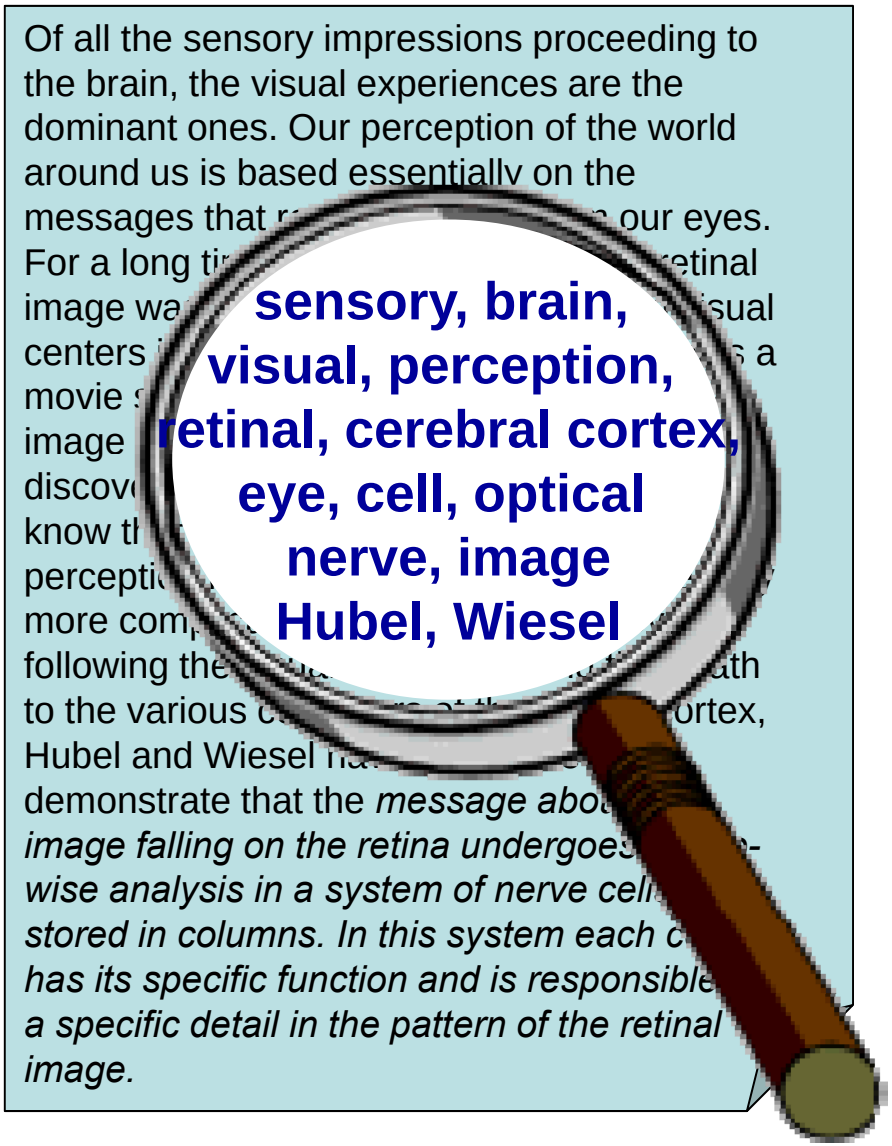
- Introduction & Research Questions
- Background & Related Works
  - Framework of Image Feature Representation
  - Generative Models for Image Features and Text
- Developed Model and Evaluation
  - Perspective Hierarchical Dirichlet Process (pHDP)
  - Evaluations
- Conclusions

Notations

- Word
  - Basic unit.
  - Item from a vocabulary indexed by {1, . . . ,V}.
- Document
  - Sequence of N words, denoted by w = (w1,w2, . . . ,wN).
- Collection
  - A total of D documents, denoted by C = {w1,w2, . . . ,wD}.
- Topic
  - Denoted by z, the total number is K.
  - Each topic has its unique word distribution p(w|z)

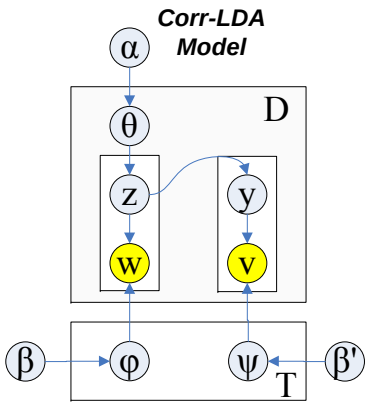
Topic Modeling - Intuitive

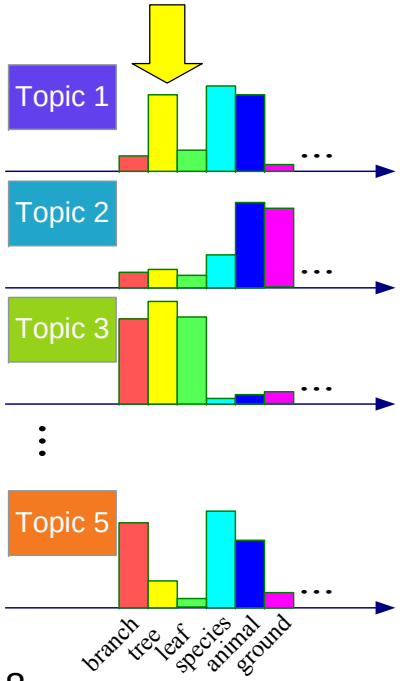
- Intuitive
  - Assume the data we see is generated by some parameterized random process.
  - Learn the parameters that best explain the data.
  - Use the model to predict (infer) new data, based on data seen so far.



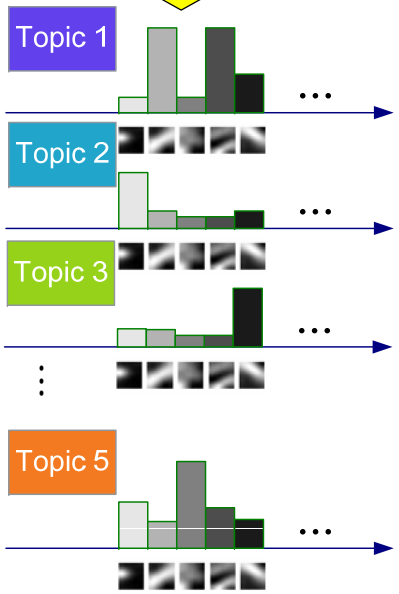
CorrLDA model (Blei, 2003) in modeling image and text

Needle-leaf forest is composed largely of straight trunked, conical tress with relatively short branches, and small, narrow, needlelike leaves. These tress are conifers. Where evergreen, the needleleaf forest provides continues and deep shade to the ground so that lower layers of vegetation are sparse or absent except for a thick carpet of mosses in many places. Species are few and large tracts of forest consist almost entirely of but one or two species.





Document-level Topic Mixture Composition



Nonparametric Hierarchical Bayesian Model

- In real-world applications, the number of semantic components in an image is not fixed, for example, a picture of clear blue sky tend to have less semantic components than an image showing crowd of people in the street.

- The Hieratical Dirichlet Process (HDP) model (Teh, 2006), is a nonparametric extension of the Latent Dirichlet Allocation (LDA)-based topic models, it enables modeling documents with countable infinite mixture components, thus provides the flexibility of modeling images whose actual semantic component numbers are unknown

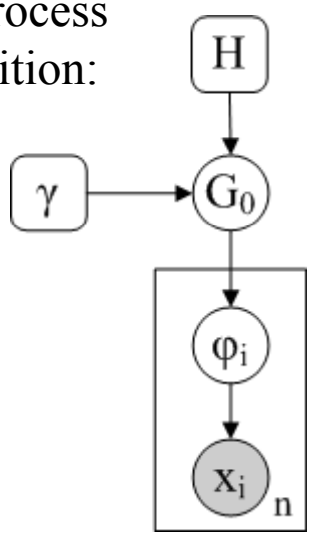
Dirichlet Process (DP) as a Non-Parametric Mixture Models

The Dirichlet Process (DP) is defined as a distribution of random probability measure $G_0 \sim \mathrm{DP}(\gamma, H)$, in which γ is a concentration parameter and H is a base measure defined on a sample space Θ. By its definition, for any finite measurable partition of Θ: $\{A_1, \ldots, A_r\}$, $(G_0(A_1), \ldots, G_0(A_r)) \sim \mathrm{Dirichlet}(\gamma H(A_1), \ldots, \gamma H(A_r))$.

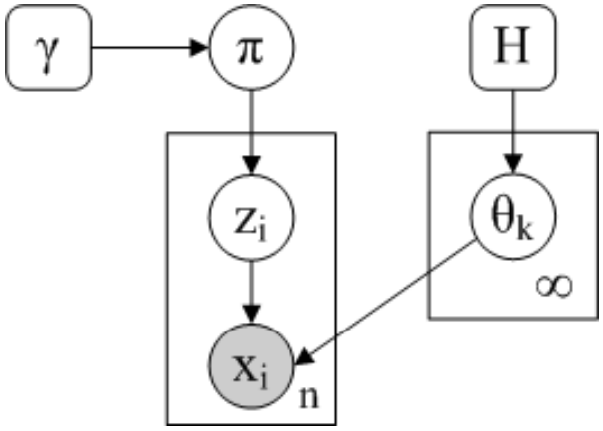
Dirichlet Process can also be constructed by stick-breaking construction as follows:

$$G_0 = \sum_{k=1}^{\infty} \beta_k \delta(\theta_k) \quad \beta_k = \alpha_k \prod_{i=1}^{k-1} (1-\alpha_i),\ \alpha_k \sim Beta(1, \gamma)$$

Dirichlet process by its definition:



Dirichlet process constructed by stick-breaking construction:



- Data sample x_i drawn from a base distribution with associated parameters Θ_k

The weights of mixture components $\boldsymbol{\beta} = \{\beta_k\}$ (k=1,…, ∞) are also refer to as $\boldsymbol{\beta} \sim \mathrm{GEM}(\gamma)$.

,in which

Hierarchical Dirichlet Process (HDP)

The Hierarchical Dirichlet Process (HDP) considers $G_0 \sim \mathrm{DP}(\gamma, H)$ as a global probability measure across the corpora and defines a set of child random probability measures $G_j \sim \mathrm{DP}(\alpha_0, G_0)$ for each document j, which leads to different document-level distribution over semantic mixture components: $(G_j(A_1),\ldots,G_j(A_r)) \sim \mathrm{Dirichlet}(\alpha_0 G_0(A_1),\ldots, \alpha_0 G_0(A_r))$

Each G_j can also be constructed by stick-breaking construction as: $G_j = \sum_{k=1}^{\infty} \pi_{jk}\delta(\theta_k)$

in whch $\boldsymbol{\pi_j}=\{\pi_{jk}\}$ (k=1,…, ∞) specifies the weights of mixture component indicator k.

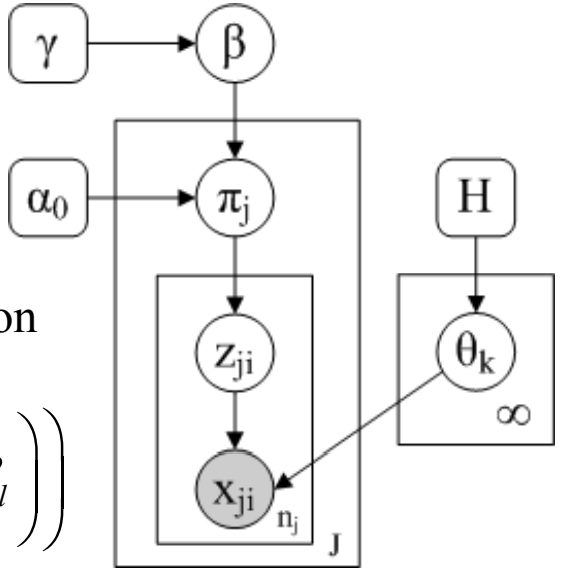
Substitute the stick-breaking construction of G_0 and G_j, it follows that:

$$\left(\sum_{k\in K_1}\pi_{jk},\ldots,\sum_{k\in K_r}\pi_{jk}\right) \sim Dirichlet(\alpha_0\sum_{k\in K_1}\beta_k,\ldots,\alpha_0\sum_{k\in K_r}\beta_k)$$

Based on the aggregation properties of Dirichlet distribution and its connection with Beta distribution, it shows that:

$$\pi_{jk} = \pi'_{jk}\prod_{l=1}^{k-1}(1-\pi'_{jl}),\ \pi'_{jk} \sim Beta\left(\alpha_0\beta_k, \alpha_0\left(1-\sum_{l=1}^{k}\beta_l\right)\right)$$

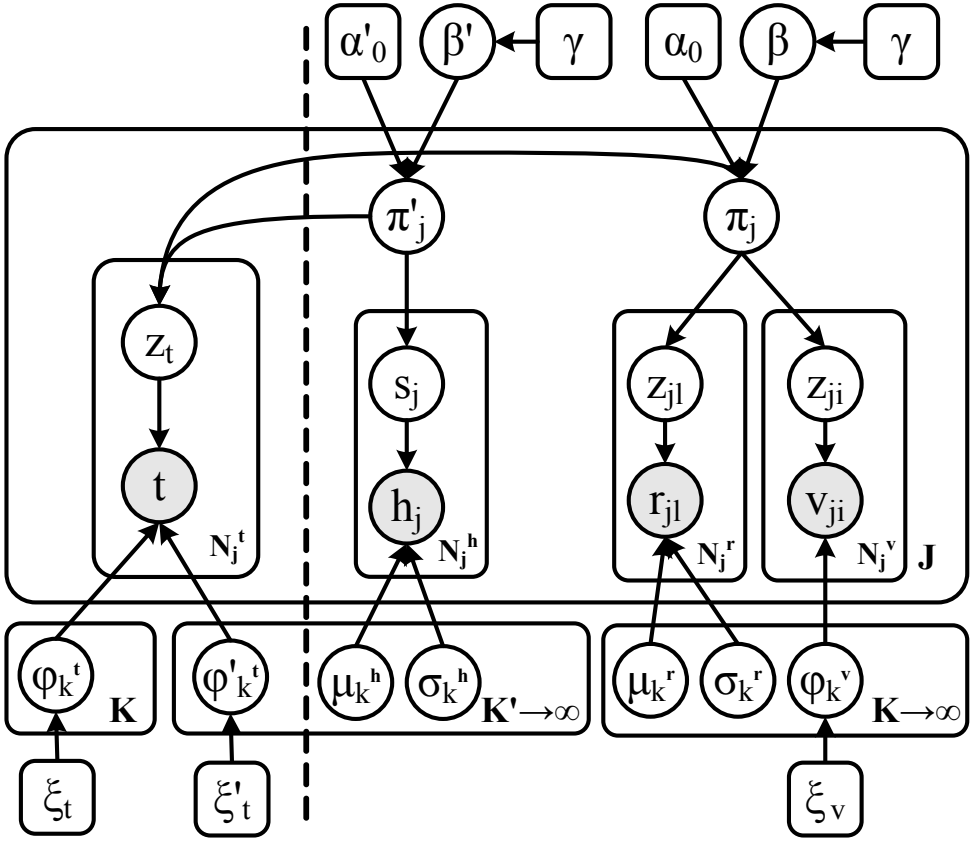
It then follows that $\boldsymbol{\pi_j} \sim \mathrm{DP}(\alpha_0, \boldsymbol{\beta})$



Stick-breaking construction of hierarchical Dirichlet process

The extended HDP model

- The HDP model can be extended to both image features and associated image tags. The extended HDP model enables us to represent image content with unbounded number of semantic components and establish correspondence between image tags and image features

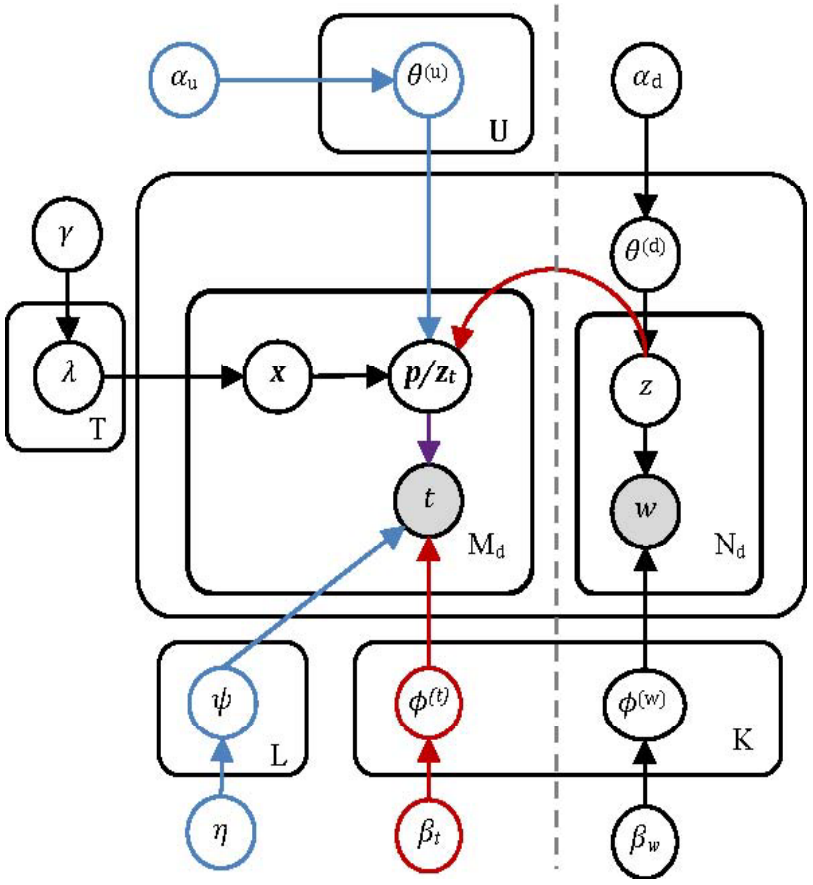


Outlines

- Introduction & Research Questions
- Background & Related Works
  - Framework of Image Feature Representation
  - Generative Models for Image Features and Text
- Developed Model and Evaluation
  - Perspective Hierarchical Dirichlet Process (pHDP)
  - Evaluations
- Conclusions

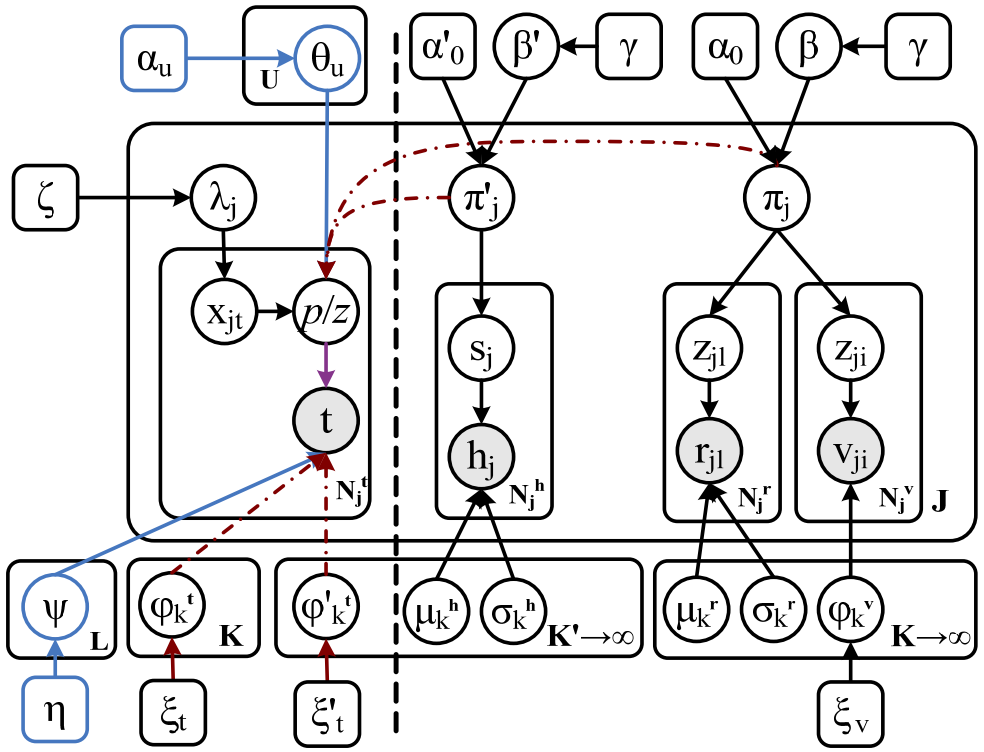
Topic-perspective (TP) model (Lu, 2010)

- The topic-perspective (TP) model is proposed to infer how both users' perspective and the resource content relate to the generation of social annotations.
- It separates the tag generation process from the generation process of the resource content. While the resource content (such as text words) is only generated from resource topics, the social tags are generated by both resource topic and user perspective.



The perspective HDP (pHDP) model

- We incorporate the users' perspective into the image tag generation process and introduce new latent variables to determine if an image tag is generated from user's perspective or from the image content.



The value of *x* (which takes values 0-2) is sampled from a multinomial distribution λ_j (with a Dirichlet prior ζ). When the value of *x* equals 0 or 1, the topical indicator of tag *t* is draw uniformly from the semantic components learned from the image contents. When *x* equals 2, a user perspective *p* will be sampled from the perspective distribution (θ_u) for user *u*, and tag *t* will be drawn from the tag distribution ψ_p of perspective *p*.

Datasets and Experimental Setup

- We investigate the performance of the proposed pHDP model and extended HDP model under an automatic image tagging experiment on the MIR-Flickr 25000 dataset (Huiskes and Lew, 2008).

- This dataset is composed of 25000 Flickr images contributed by a total of 9862 Flickr users and covers a wide spectrum of image categories. In the collection, there are a total of 64037 unique tags; and the average number of tags per image is 8.94. In our image tagging experiment, we use a 50% subset of the MIR-Flickr collection as training data and the other 50% as testing data (with tags removed). On constructing the two subsets, we ensure that tagged images from the same user are equally split to both subsets.

Model Estimation and Illustration

- The estimation of the proposed models is achieved by performing Gibbs sampling on the training dataset until convergence .Once the model estimation is finished, we will be able to visualize the uncovered semantic components and user perspectives.

Top tags	Probability
indoor	0.0871
people	0.0852
female	0.050
male	0.0447
portrait	0.0345
selfportrait	0.0201
bw	0.0154
blackandwhite	0.0092
night	0.0072

Top tags	Probability
indoor	0.1015
structures	0.0483
chair	0.0194
book	0.0145
mug	0.0112
cushion	0.0097
casecubiche	0.0065
mushroom	0.0059
cup	0.0048

Top tags	Probability
clouds	0.0363
lake	0.0331
sky	0.0297
sunset	0.0231
iceland	0.0198
pink	0.0145
blue	0.0132
eco	0.0099
explore	0.0048

Subset of uncovered semantic components

Model Estimation and Illustration (continue)

- Semantic components are derived from image features and indicate the visual contents in images. The user perspectives , on the other hand, show user's preferences and subjective feelings during image tagging.

Top tags	Probability
baby	0.0894
love	0.0671
pie	0.0447
heart	0.0335
ring	0.0279
handbag	0.0224
gift	0.0112
sis	0.0097
top	0.0056

Top tags	Probability
beautiful	0.0773
joy	0.0483
argentina	0.0263
hair	0.0198
smile	0.0193
myself	0.0145
happy	0.0132
rare	0.0097
cute	0.0066

Top tags	Probability
nature	0.0654
nikond200	0.0577
autumn	0.0538
impressedbeauty	0.0462
spring	0.0423
leaves	0.0385
anawesomeshot	0.0308
supershot	0.0231
naturesfinest	0.0193

Subset of uncovered user perspectives

Image Tagging Experiment

- We calculate the probability of tagging an image j from user u with different tags. Tags with highest probability are used for tagging. After that, the predicted top-ranked image tags are compared with the ground truth for validation.
- If a predicted tag finds exact match in the ground truth tags, it will be considered as one hit.

User ID: 17875539@N00
Title: City with Ice
Tags: toronto, torontoharbour, canada, ontario, clouds, lake, night, sky, structures, water, ice, winter, snow, frozen, city, cityscape, nikon, nikond200, cold, landscape, lake, water, outdoor, outdoorphotography, frost, frosty

Top ranked tag	Probability
canada	**0.0606**
ontario	**0.0597**
water	0.0523
sky	0.0474
lake	0.0427
toronto	**0.0322**
clouds	0.0295
outdoor	0.0257
structures	0.0232

tags predicted by pHDP

Top ranked tag	Probability
structures	0.0538
sky	0.0329
night	0.0265
clouds	0.0159
water	0.0096
sunset	0.0075
sea	0.0074
buildings	0.0053
snow	0.0035

tags predicted by extended HDP

Image Tagging Experiment (continue)

- When tagging a new image from the same user, the pHDP model will smooth the document-level predictive tag distribution with user's perspective and allow for tagging with *location* tags ('ontario', 'canada') and *topic* tags (such as 'clouds', 'lake', 'night', sky and 'water').

User ID: 17875539@N00
Title: Some say in ice
Tags: lake, night, ontario, toronto, torontoharbour, canada, ice, frozen, winter, dawn, morning, nature, sky, water, landscape, snow, nikon, nikond200, cold, cherrybeach, outdoor, natural, frost, frosty

Top ranked tag	Probability
sky	0.0458
canada	**0.0446**
ontario	**0.0439**
structures	0.0395
clouds	0.0244
water	0.0226
toronto	**0.0209**
sunset	0.0202
sea	0.0195

tags predicted by pHDP

Top ranked tag	Probability
sky	0.0375
structures	0.0306
clouds	0.0300
water	0.0199
sunset	0.0173
sea	0.0167
flower	0.0162
transport	0.0139
outdoor	0.0124

tags predicted by extended HDP

Image Tagging Experiment (continue)

- Other user contextual information is also captured in user's perspectives. Thus the pHDP model succeeds in tagging image with both *location* tags (such as 'malaysia') and *type* tags (camera settings, like 'nikon'). Tags predicted by the pHDP model also involve *subjective* tag, like 'interestingness'.

User ID: 9352758@N04
Title: Spread your wings and and fly away
Tags: female, people, portrait, tree, children, play, fun, windy, wind, fly, kid, girl, hair, malaysia, landscape, nikon, d50, movement, vacation, travel, diamond class photographer, an awesome shot

Top ranked tag	Probability
people	0.0631
female	0.0346
portrait	0.0305
flower	0.0239
outdoor	0.0427
nikon	0.0238
malaysia	**0.0227**
sky	0.0208
interestingness	0.0207

tags predicted by pHDP

Top ranked tag	Probability
people	0.0518
female	0.0290
portrait	0.0283
sky	0.0242
outdoor	0.0219
clouds	0.0213
tree	0.0187
flower	0.0156
water	0.0134

tags predicted by extended HDP

Outlines

- Introduction & Research Questions
- Background & Related Works
  - Framework of Image Feature Representation
  - Generative Models for Image Features and Text
- Developed Model and Evaluation
  - Perspective Hierarchical Dirichlet Process (pHDP)
  - Evaluations
- Conclusions

Automatic Image Tagging and Evaluation

- The prediction of image tags for the testing images is achieved by performing another Gibbs sampling on testing images to estimate the document-level distribution of switch variable and semantic components, with a fixed set of semantic components and user perspectives estimated from the training dataset. On the convergence of Gibbs sampling, the probability of tagging an image *j* from user *u* with tag t_j is:

$$p(t_j) = p(x_{jt} = 0,1)\sum_{k=1}^{K} p(t_j \mid z_k)\, p_{test}(z_k \mid j) +$$

$$p(x_{jt} = 2)\sum_{l=1}^{L} p(t_j \mid p_l)\, p_{test}(p_l \mid u)$$

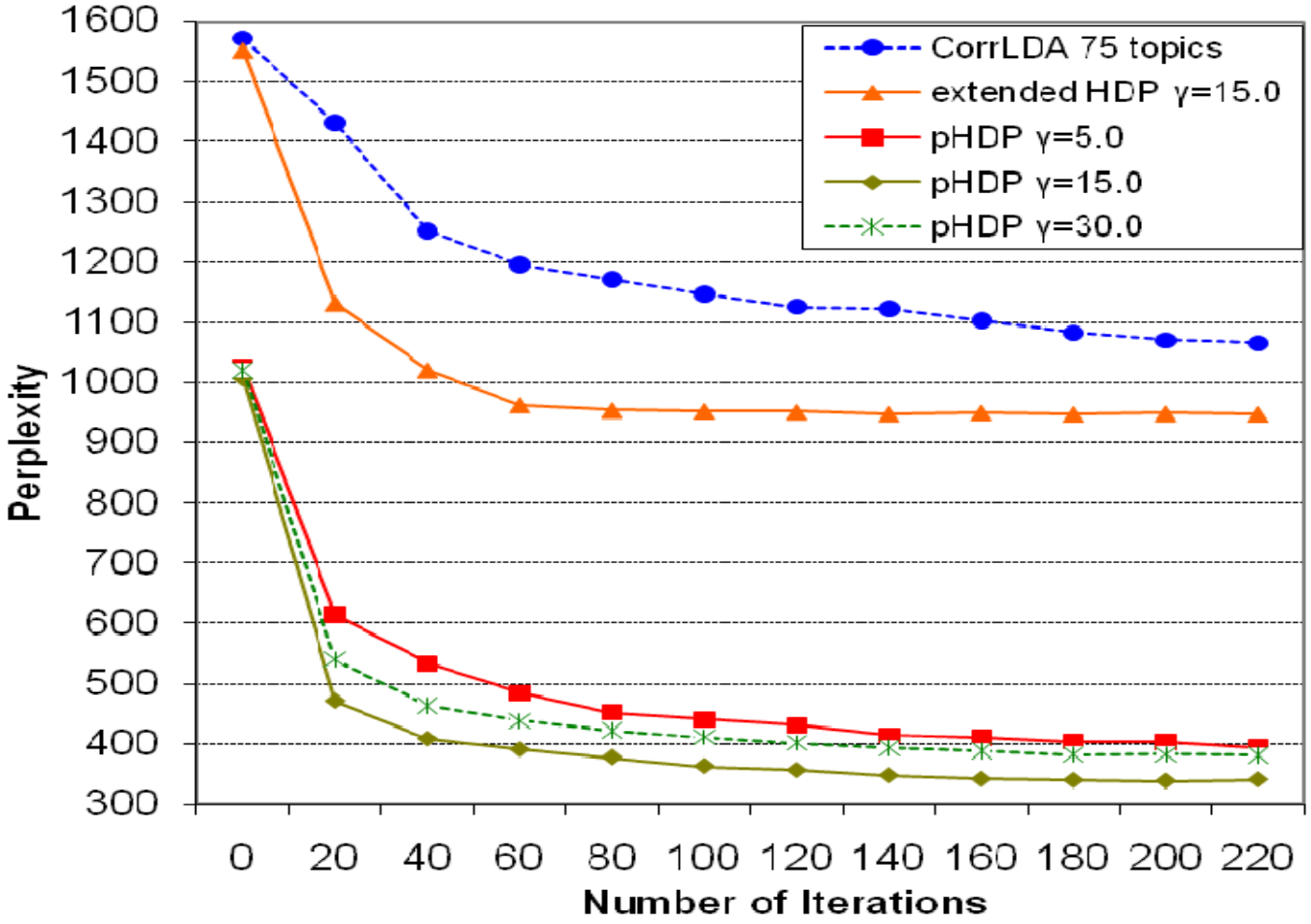
Tag Perplexity Comparison

- The perplexity is a standard criterion for generative probabilistic models that evaluates how well the model predicts the testing data. The perplexity of a testing image dataset D_{test} is:

$$perplexity(D_{test}) = \exp\left[\frac{-\sum_{j=1}^{D_{test}} \log(p(\mathbf{t_j}))}{\sum_{j=1}^{D_{test}} N_j^t}\right]$$

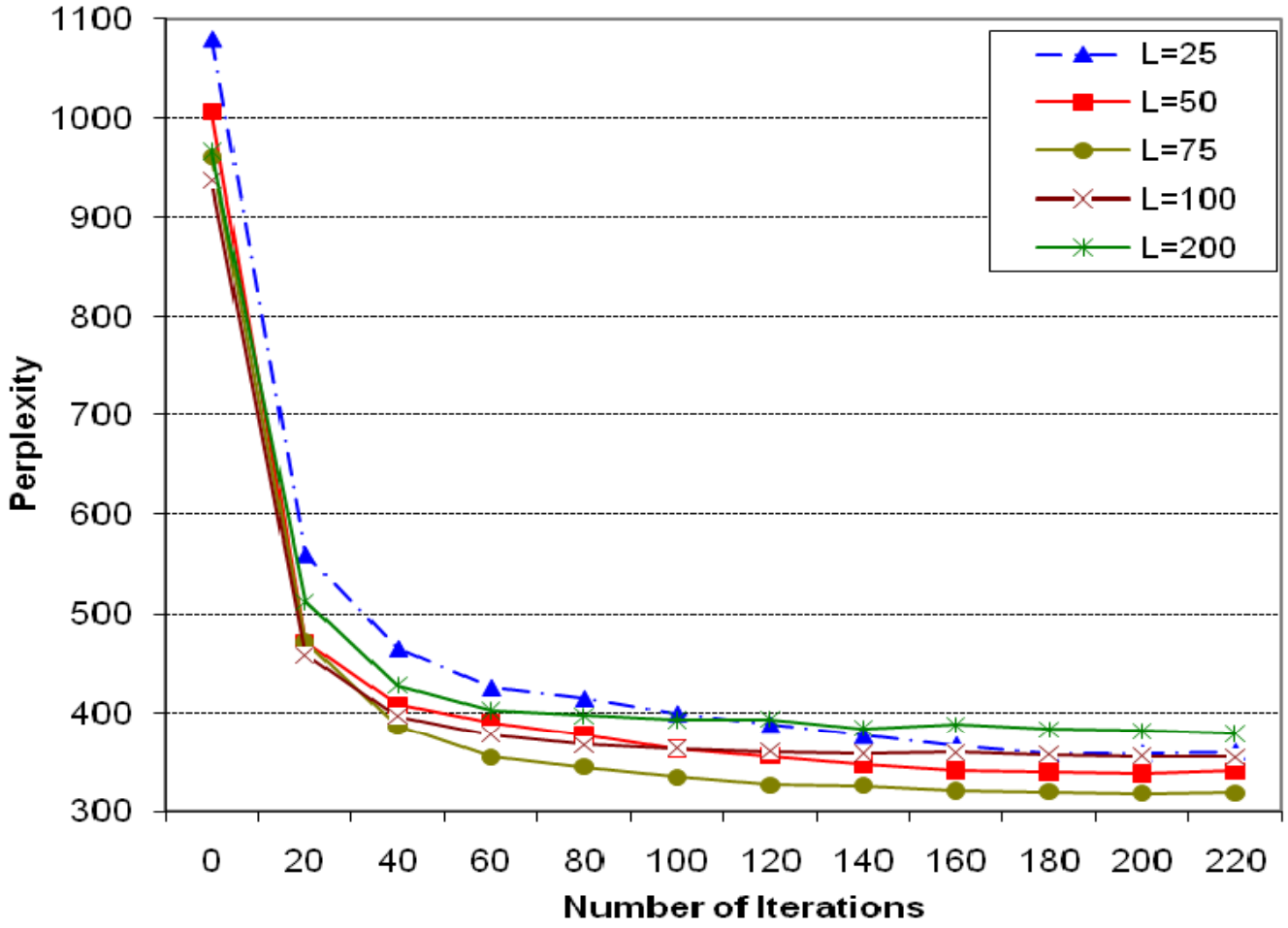
- The perplexity score for a model is the lower the better.

The perplexity comparison of the proposed models and baseline Corr-LDA model



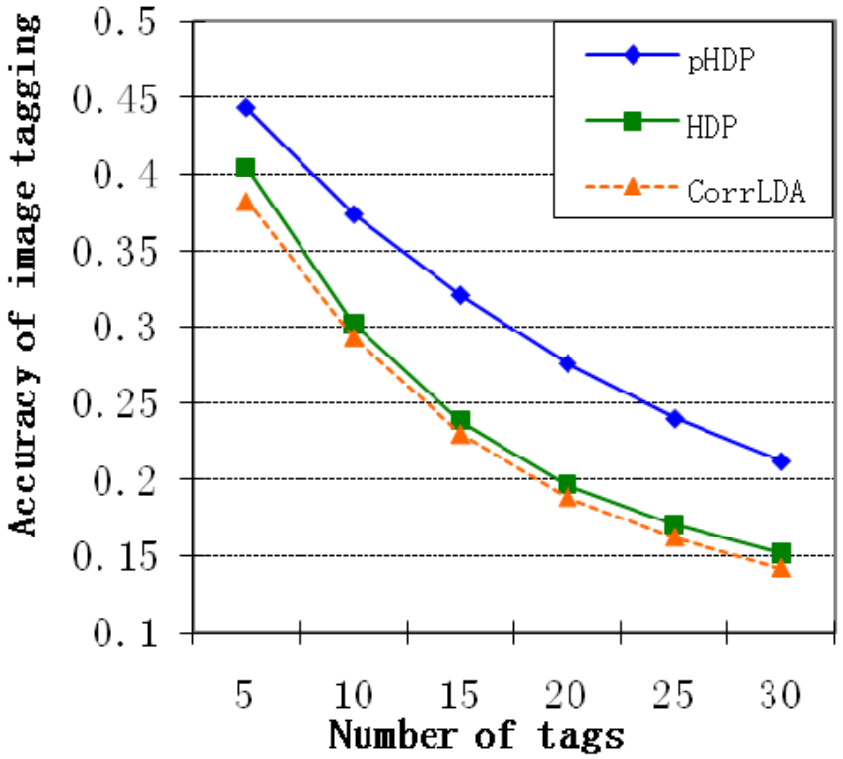
Perplexity changing over iterations

Perplexity of pHDP model (with different perspective number L)



Comparison of average image tagging accuracy

- The pHDP model, as it integrates the user perspective information, significantly outperforms both CorrLDA model and extended HDP model in predicting image tags for different users.



Outlines

- Introduction & Research Questions
- Background & Related Works
  - Framework of Image Feature Representation
  - Generative Models for Image Features and Text
- Developed Model and Evaluation
  - Perspective Hierarchical Dirichlet Process (pHDP)
  - Evaluations
- Conclusions

Conclusions

- The contribution is twofold. Firstly, we extend the HDP model to both image features and associated image tags. Secondly, we incorporate the user's perspectives into the image tag generation process and introduce new latent variables to determine if an image tag is generated from user's perspectives or from the image content.

- Based on the proposed pHDP model, we achieve automatic image tagging with users' perspective. Experimental results show that the proposed model not only generates useful information about semantic components and user perspectives from tagged images, but also achieves better performance in the task of automatic image tagging compared to CorrLDA model and extended HDP model.

Reference

[1] D.M. Blei, and M.I. Jordan, Modeling annotated data The 26th annual international ACM SIGIR conference on Research and development in informaion retrieval, ACM, Toronto, Canada, 2003, pp. 127-134.

[2] Henderson, J.M. and Hollingworth, A. High level scene perception. Annual Review of Psychology, 50:243–271, 1999.

[3] C. Siagian and L. Itti, Rapid Biologically-Inspired Scene Classification Using Features Shared with Visual Attention, IEEE TPAMI, pp. 300-312, 2007.

[4] A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, Content-based image rerieval at the end of the early years, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 22, no. 12, pp. 1349–1380, 2000.

[5] Y. Teh, M. Jordan, M. Beal, and D. Blei. Hierarchical Dirichlet process. Journal of the American Statistical Association, 101(476):1566-1581, 2006

[6] K. Bischoff, C.S. Firan, W. Nejdl, and R. Paiu, Can All Tags be Used for Search?, CIKM'08, Napa Valley, California, USA, 2008, pp. 203-212.

[7] S. Sen, S.K.T. Lam, A.M. Rashid, D. Cosley, D. Frankowski, J. Osterhouse, F.M. Harper, and J. Riedl, Tagging, communities, vocabulary, evolution, CSCW'06, Banff, Alberta, Canada, 2006.

[8] Amr Ahmed, Eric P. Xing, William W. Cohen, Robert F. Murphy, Structured Correspondence topic models for mining captioned figures in biomedical literature, Proceedings of the 15th ACM SIGKDD International conference on Knowledge discovery and data mining, June 28-July 01, 2009, Paris, France.

[9] X. Chen, C. Lu, Y. An, and P. Achananuparp. Probabilistic Models for Topic Learning from Images and Captions in Online Biomedical Literatures. In the Proceedings of 18th ACM Conference on Information and Knowledge Management (CIKM'09)

[10] D. Zhou, J. Bian, S. Zheng, H. Zha, and C.L. Giles, Exploring Social Annotations for Information Retrieval, WWW 2008, Beijing, China, 2008, pp. 715-724.

[11] C. Lu, X. Hu, X. Chen and J. Park. The topic-perspective model for social tagging systems, The 16th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD'10), July 25-28, 2010, Washington D.C., USA. Pp. 683-692.

[12] Sivic, J., Zisserman, A.: Video Google: A Text Retrieval Approach to Object Matching in Videos. International Conference on Computer Vision. (2003) 1470– 1477

[13] J. Matas, O. Chum, U. M., T. Pajdla. Robust wide baseline stereo from maximally stable extremal regions. In BMVC, 2002.

Questions?

Backup Slides