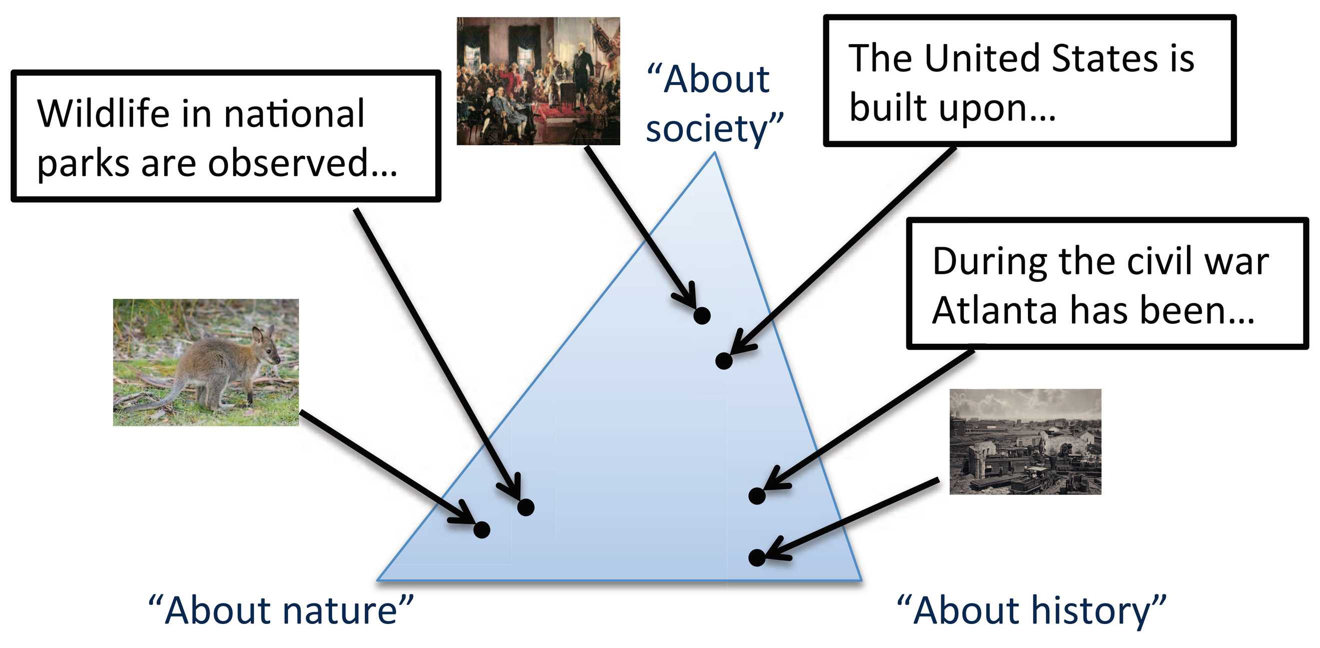
Learning Cross-modality Similarity for Multinomial Data

Yangqing Jia[1] Mathieu Salzmann[2] Trevor Darrell[1]

[1]EECS & ICSI, UC Berkeley [2]TTI Chicago

Summary

- Many applications involve multiple modalities.
- We learn a latent topic space that models the joint semantics of multiple modalities, such as images and loosely related narrative text.



Existing Methods

Multiple Kernel Learning

- Combine kernels from multiple modalities (e.g., Lanckriet et al. JMLR'04).
- These methods are discriminative and do not learn cross-modality transfers.

Canonical Correlation Analysis

- Find projection directions on which multiple modalities are maximally correlated.
- May not work well when data follow non-Gaussian, sparse distributions.

Shared Latent Variable Models

- Designed for dense, real-valued feature spaces (e.g. GPLVM).
- Effective in applications such as human pose estimation (Ek PhD Thesis'09, Salzmann et al. AISTATS'10), image synthesis (Shon et al. NIPS 05), and domain transfer (Saenko et al. ECCV'10).
- Not suitable for data following multinomial distributions.

Latent Topic Models

- Based on the LDA model, assuming that words correspond to real-world objects.
- Aims to find correspondence between words and local image patches (e.g., Barnard et al. JMLR'03, Blei et al. SIGIR'03, Wang et al. CVPR'09).
- Requires each "document" to contain both/all modalities.
- Modalities are not symmetric - the model has a main modality (usually images) and dependent modalities.
- Fail to use loose text descriptions containing abstract words.

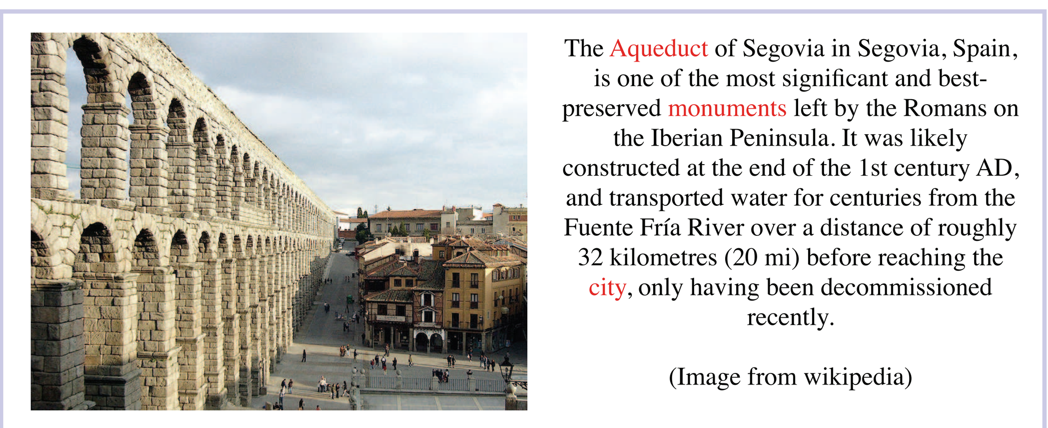


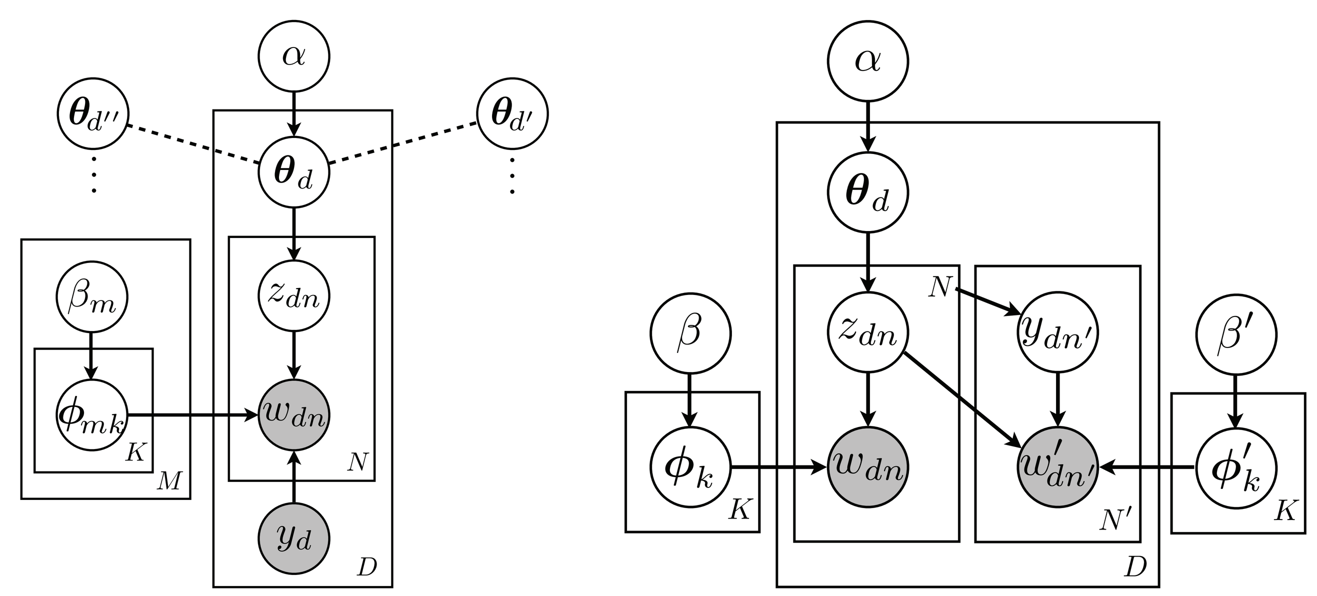
Figure: words corresponding to real objects (the red ones) in the picture are usually scarce, and loose descriptions may contain much richer information.

Our Approach

- Learn cross-modal topics from uni-modal documents.
  - Treats each modality equally, and naturally extends to more than 2 views.
- Cross-modality similarity are introduced in the document level.
  - Learns cross-modal semantic information.
  - Cross-modal inquiry becomes simple distance comparison.
- The topic proportions in each document can be viewed as a common latent representation for all modalities.

Multi-modal Document Random Field

Graphical models of our method and Corr-LDA:



(a) MDRF (b) Corr-LDA

- We define a Markov random field on the document level with potential functions:

$$\psi(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j) = \exp(-\lambda f(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j))$$

where $f(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j) = \frac{1}{2}(D(\boldsymbol{\theta}_i \| \boldsymbol{\theta}_j) + D(\boldsymbol{\theta}_j \| \boldsymbol{\theta}_i))$

No strict correspondence needed for the data, can handle flexible similarity supervision (documents within the same modality, or from different modalities).

Generative Procedure

- For each topic k , sample the word distributions $\phi_{mk} \sim \text{Dir}(\phi|\beta_m)$ for each modality m.
- Sample the D topic proportions $\boldsymbol{\theta}_{1\ldots D}$ from the distribution

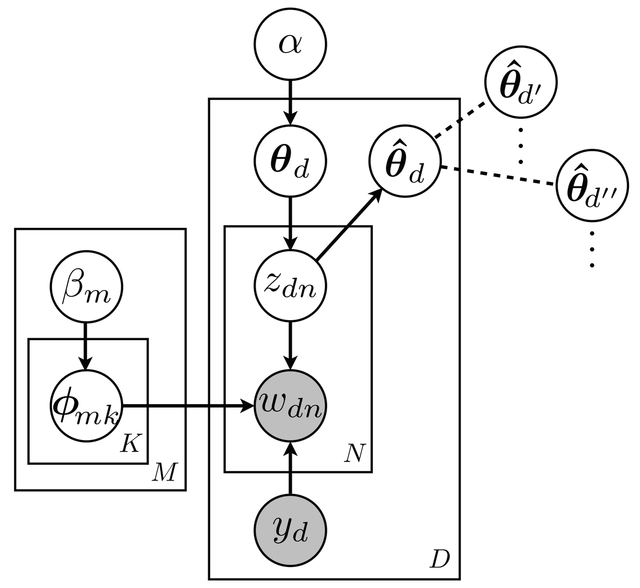
$$p(\boldsymbol{\theta}_{1\ldots D}|\alpha, \mathcal{G}) = \frac{1}{Z}\exp(-\lambda \sum_{i,j\in\mathcal{E}} f(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j)) \prod_{d=1}^{D} \text{Dir}(\boldsymbol{\theta}_d|\alpha),$$

- For each document d, sample its modality y_d from a uniform distribution over $\{1, \cdots, M\}$.
- For each word w_{dn}:
  - Sample a topic $z_{dn} \sim \text{Multi}(z|\boldsymbol{\theta}_d)$, and sample a word $w_{dn} \sim \text{Multi}(w|\boldsymbol{\phi}_{y_d z_{dn}})$.

Collapsed Gibbs Sampling

- Solving the original model is difficult due to the coupled $\boldsymbol{\theta}$.
- We solve an empirical MDRF model via Gibbs sampling.
- Sampling the topic z of a word in document d by collapsing $\boldsymbol{\theta}$ and $\boldsymbol{\Phi}$:

$$P(z=k|\mathcal{D}, \mathbf{z}_{-w}, \alpha, \beta) \propto \frac{n_{dk}^{(d)} + \alpha}{\sum_{k=1}^{K} n_{dk}^{(d)} + K\alpha} \times \frac{n_{kw}^{(m)} + \beta_y}{\sum_{w=1}^{V_m} n_{kw}^{(m)} + V_m\beta_m} \times \prod_{d', (d,d')\in\mathcal{E}} \exp\left(\lambda f(\hat{\boldsymbol{\theta}}_{d,-z}, \hat{\boldsymbol{\theta}}_{d'}) - \lambda f(\hat{\boldsymbol{\theta}}_{d,z=k}, \hat{\boldsymbol{\theta}}_{d'})\right),$$



Experiments

Dataset

- We collected the Wikipedia "Picture of the Day" dataset:
  - `http://www.eecs.berkeley.edu/~jiayq/wikipedia_potd/`
- Images and loose text descriptions from Nov 2004 to Oct 2010.
  - Bag-of-words model for both image and text.

Protocol

- Image topic distributions θ_i are inferred for each testing image.
- For each text query $\mathbf{w} = \{w_1, w_2, \cdots, w_N\}$, return a sorted list of images based on the score

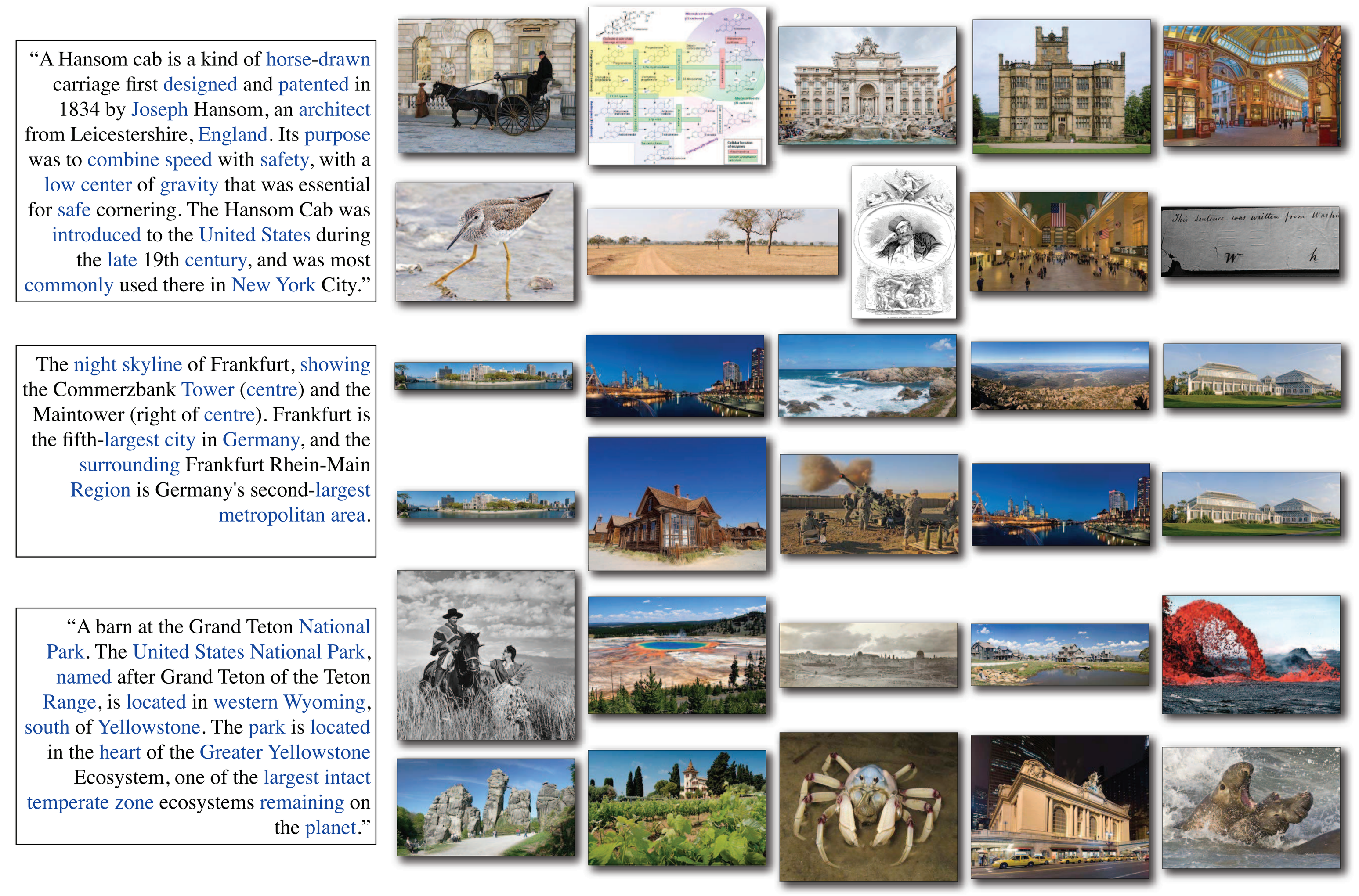
$$s_i = p(\mathbf{w}|\boldsymbol{\theta}_i) = \prod_{n=1}^{N} p(w_n|\boldsymbol{\theta}_i)$$

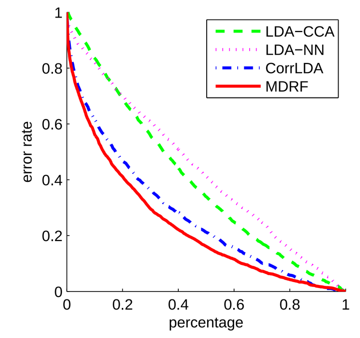
$p(w_n|\boldsymbol{\theta}_i)$ can be pre-computed for each image - no marginalization needed during retrieval time.

- Evaluation criterion: error rate (whether ground-truth has been retrieved or not) vs. percent of total test images returned.
- Baseline: LDA + Nearest Neighbor, LDA + CCA, Corr-LDA.

Retrieval Results

(For each query, the top row comes from MDRF and the bottom row from Corr-LDA)





Method	AUC value
LDA-NN	43.15 ± 1.95
LDA-CCA	39.44 ± 2.27
Corr-LDA	26.94 ± 1.87
MDRF	$\mathbf{23.14 \pm 1.49}$

Future Directions

- Non-parametric approaches to determine the number of topics.
- Factorized latent topic spaces for images and text.
- Online large-scale algorithms for cross-modal information transfer.