Learning Cross-modality Similarity for Multinomial Data
Yangqing Jia\textsuperscript{1} \hspace{1cm} Mathieu Salzmann\textsuperscript{2} \hspace{1cm} Trevor Darrell\textsuperscript{1}
\textsuperscript{1}EECS & ICSI, UC Berkeley \hspace{1cm} \textsuperscript{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

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

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 is 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:

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 (usual images) and dependent modalities.
- Fail to use loose text descriptions containing abstract words.

Methods

- Combine kernels from multiple modalities (e.g., Landrieu 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.
- Not suitable for data following multinomial 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 ’99, Salzmann et al. AGTAS’10), image synthesis (Shon et al. NIPS ’06), and domain transfer (Saenko et al. ECCV’10).
- Not suitable for data following multinomial distributions.

Experiments

- We collected the Wikipedia “Picture of the Day” dataset:
  - \url{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 $\phi_i$ are inferred for each testing image.
- For each test query $q = \{w_1, w_2, \ldots, w_N\}$, return a sorted list of images based on the score

$$s_i = p(w_q | \theta) = \sum_j p(w_q | \phi_j)$$

- Evaluation criterion: error rate (whether ground-truth has been retrieved or not) vs. percentage 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.

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.