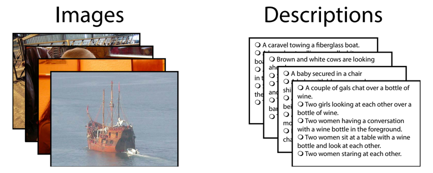
A Probabilistic Framework for Multimodal Retrieval using Integrative Indian Buffet Process

Bahadir Ozdemir[1], Larry S. Davis[2]

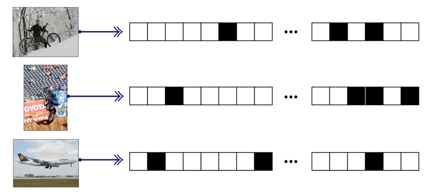
1: Department of Computer Science, University of Maryland, College Park, MD, USA
2: Institute for Advanced Computer Studies, University of Maryland, College Park, MD, USA

Unsupervised Image Retrieval

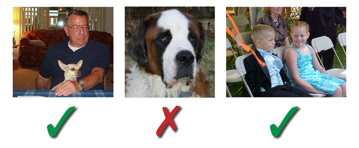
- Multimodal retrieval
  - Bridging the semantic gap
    - Text & image integration



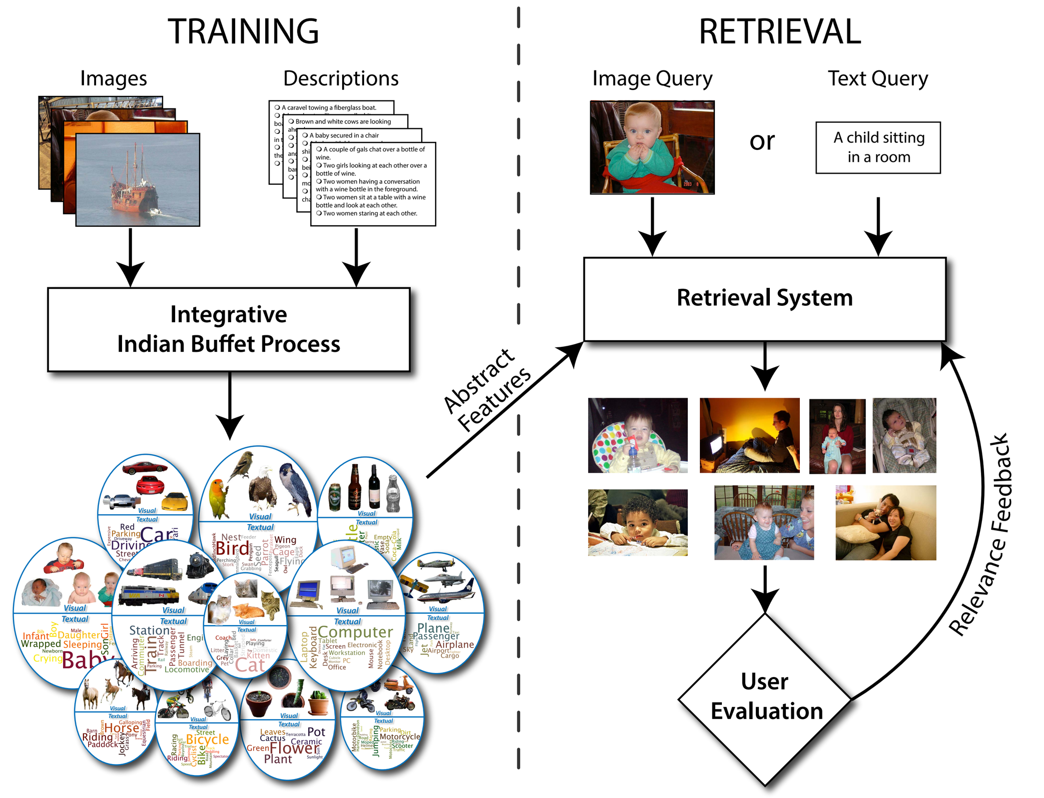
- Efficient query processing
  - Encoding into compact binary strings
    - Efficiency in search & indexing
    - Storage capacity



- Query refinement
  - Query-specific relevance judgments
  - User preference learning

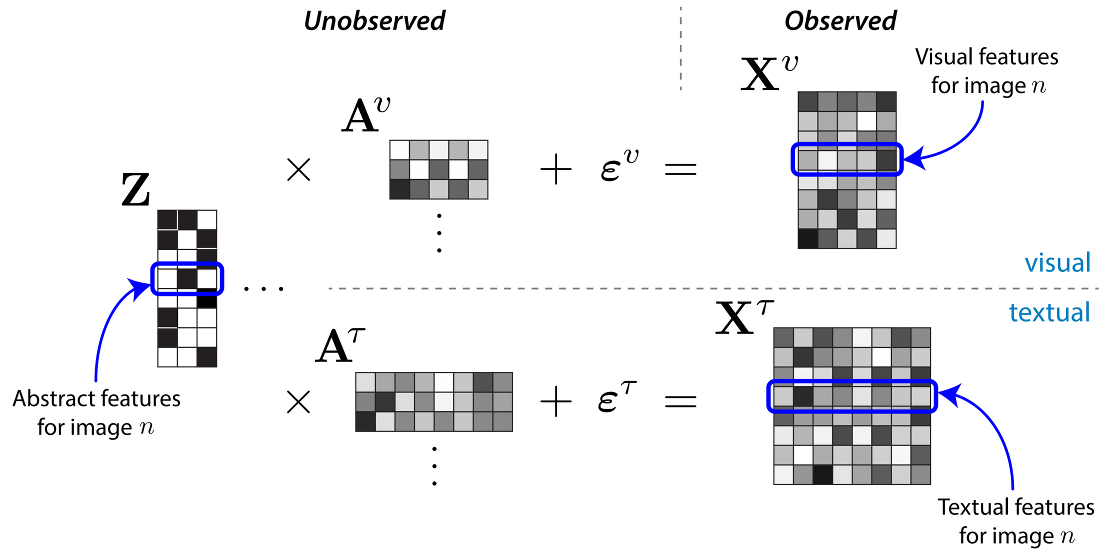


Schematic Overview of the iIBP Algorithm



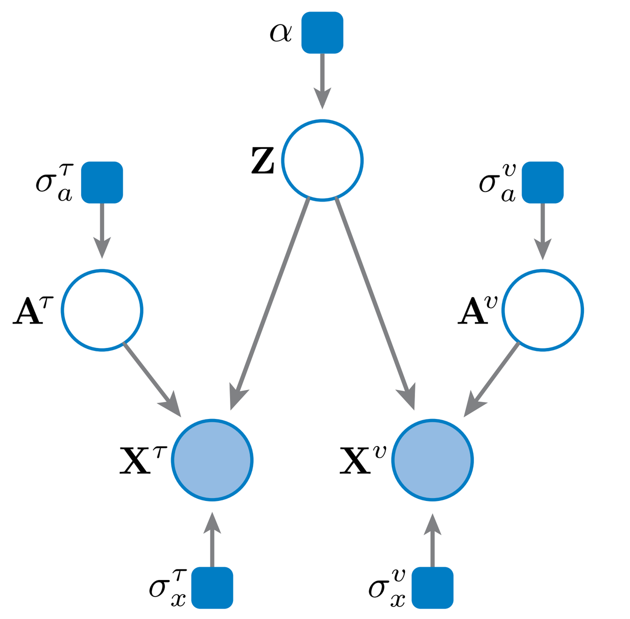
The Latent Abstract Feature Model

Visual data $\mathbf{X}^v$ is a product of $\mathbf{Z}$ and $\mathbf{A}^v$ with some noise; and similarly the textual data $\mathbf{X}^\tau$ is a product of $\mathbf{Z}$ and $\mathbf{A}^\tau$ with some noise.



Graphical Model for the Integrative IBP

Considering the infinite limit of a distribution on finite binary matrices.



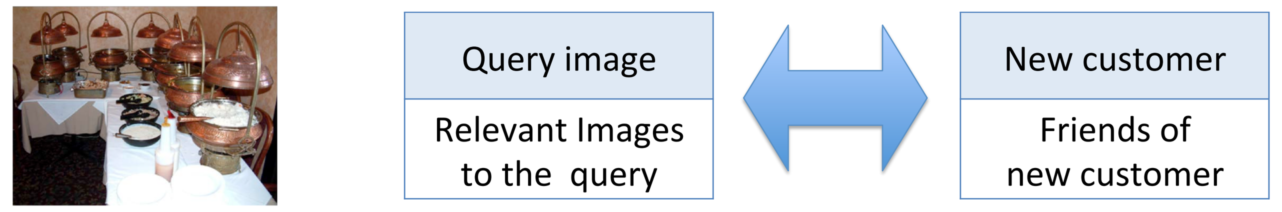
$$\mathbf{Z} \sim \mathrm{IBP}(\alpha)$$

$$\mathbf{a}_k^* \sim \mathcal{N}(\mathbf{0}, (\sigma_a^*)^2\mathbf{I})$$

$$\mathbf{x}_n^* \,|\, \mathbf{z}_n, \mathbf{A}^* \sim \mathcal{N}(\mathbf{z}_n\mathbf{A}^*, (\sigma_x^*)^2\mathbf{I})$$

Query Extension to IBP

Based on the same Indian buffet analogy [1]:
New customer (query) and his friends have similar sense of taste.

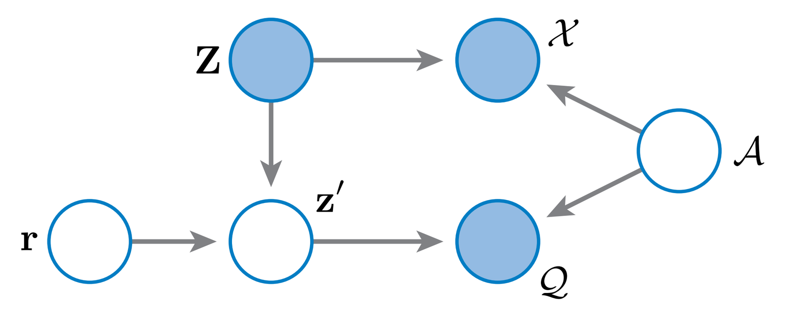


- Sampling dishes in proportion to their
  - **popularity** among **friends** and
  - **unpopularity** among **non-friends**.

$$\frac{\#\textbf{friends}\text{ who sampled that dish} + \#\textbf{non-friends}\text{ who did }\textbf{not}\text{ sampled that dish}}{\#\text{customers}}$$

- Friendship status of the nth customer: $r_n \sim \mathrm{Bernoulli}(\theta)$

Graphical Model for the Retrieval Model



- Fix the number of abstract features, K
- Sort images w.r.t. $\mathbb{E}[r_n|\mathcal{Q}, \mathbf{Z}, \mathcal{X}]$.
  - Computing the exact expectation requires 2^K computations.
- Monte Carlo approximation to $\mathbf{r}$ by resampling $\mathbf{z}'$ several times.

Inference by MCMC for the Retrieval Model

- Collapsed Gibbs sampling: Integrate out $\mathbf{r}$ and $\mathcal{A}$; sample $\mathbf{z}'$ for J times from

$$p(z'_k = 1|\mathbf{z}'_{-k}, \mathcal{Q}, \mathbf{Z}, \mathcal{X}) \propto p(z'_k = 1|\mathbf{Z})\, p(\mathcal{Q}|\mathbf{z}', \mathbf{Z}, \mathcal{X}).$$

  - Predictive probability is

$$p(z'_k = 1|\mathbf{Z}) = \frac{\theta m_k + (1-\theta)(N - m_k) + \frac{\alpha}{K}}{N + 1 + \frac{\alpha}{K}}$$

  - and collapsed likelihood function is similar to the previous one.
- Monte Carlo approximation, conditioned on $\mathbf{z}'^{(1)}, \ldots, \mathbf{z}'^{(J)}$, is

$$\hat{\mathbb{E}}[r_n|\mathcal{Q}, \mathbf{Z}, \mathcal{X}] = \frac{1}{J}\sum_{j=1}^{J} p(r_n = 1|\mathbf{z}'^{(j)}, \mathbf{Z})$$

Feedback Extension to IBP

The new customer has imaginary friends who reflect his preferences (bias).

- Sampling probability of a dish:

$$\frac{\#\text{friends who sampled that dish} + \#\text{non-friends who did not sampled that dish} + \#\textbf{imaginary friends who sampled that dish}}{\#\text{customers} + \#\textbf{imaginary friends}}$$

- Number of imaginary friends:

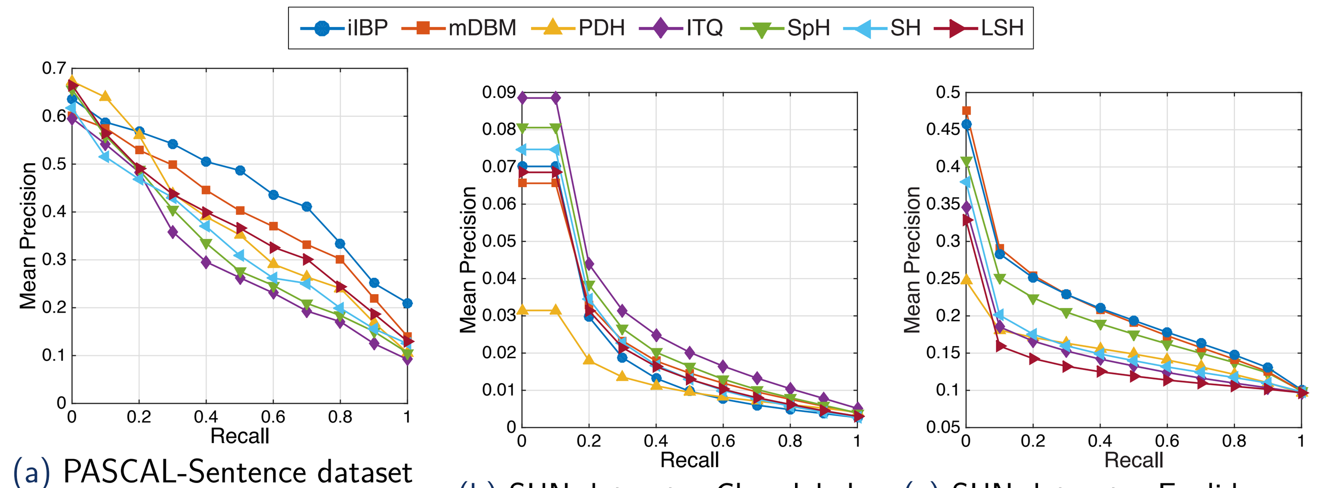
$$N_u \sim \mathrm{Binomial}(\gamma, N)$$

- Number of imaginary friends who sampled the kth dish: $0, 1, 2, \ldots, N_u$

$$m_{u,k} \,|\, N_u \sim \mathrm{Binomial}(\phi, N_u)$$

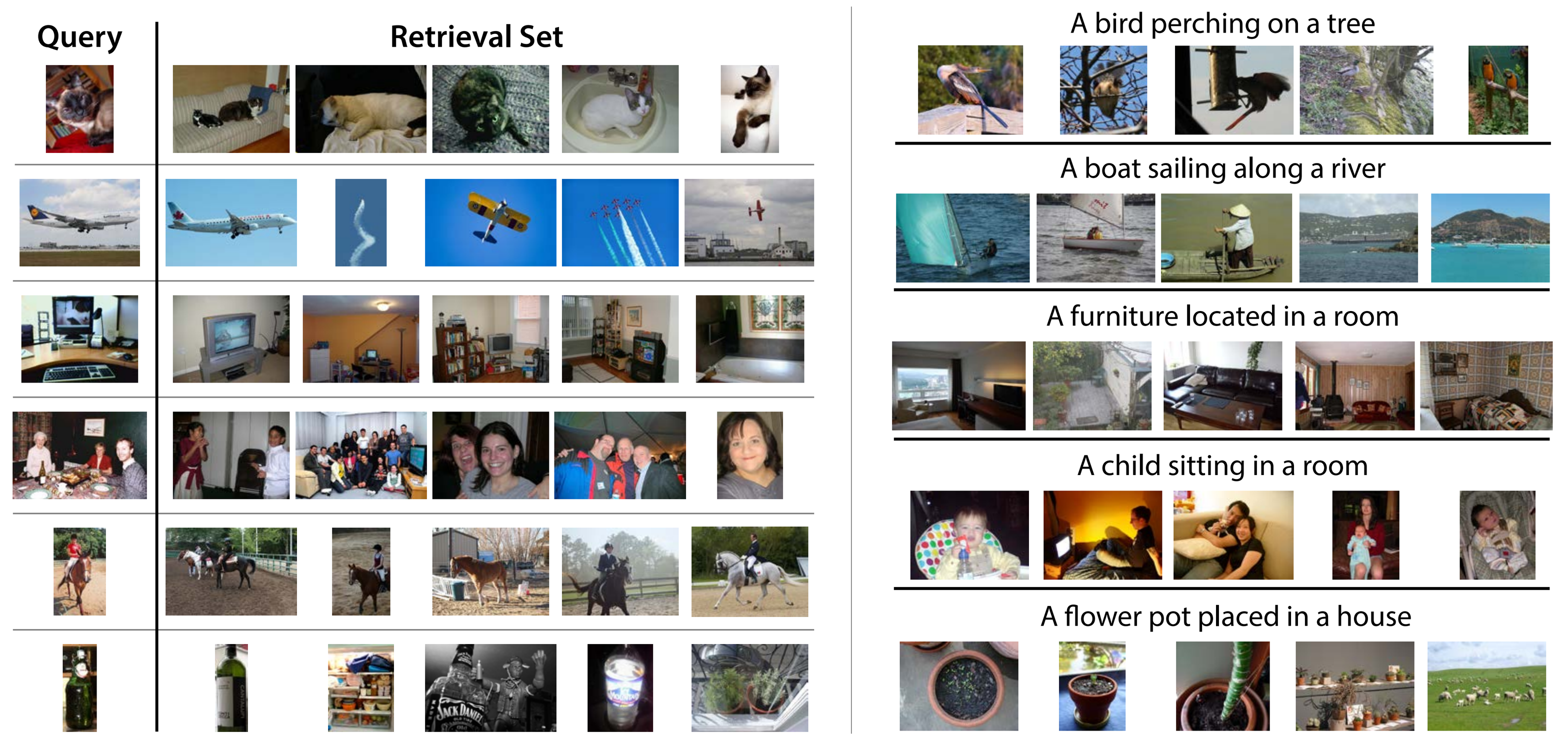
Retrieval Experiments – Quantitative Analysis

- The result of category retrieval for all query types:
  - image-to-image and text-to-image queries.
- Our method (iIBP) is compared with the state-of-the-art methods on the PASCAL-Sentence dataset[2] and SUN-Attribute dataset[3].

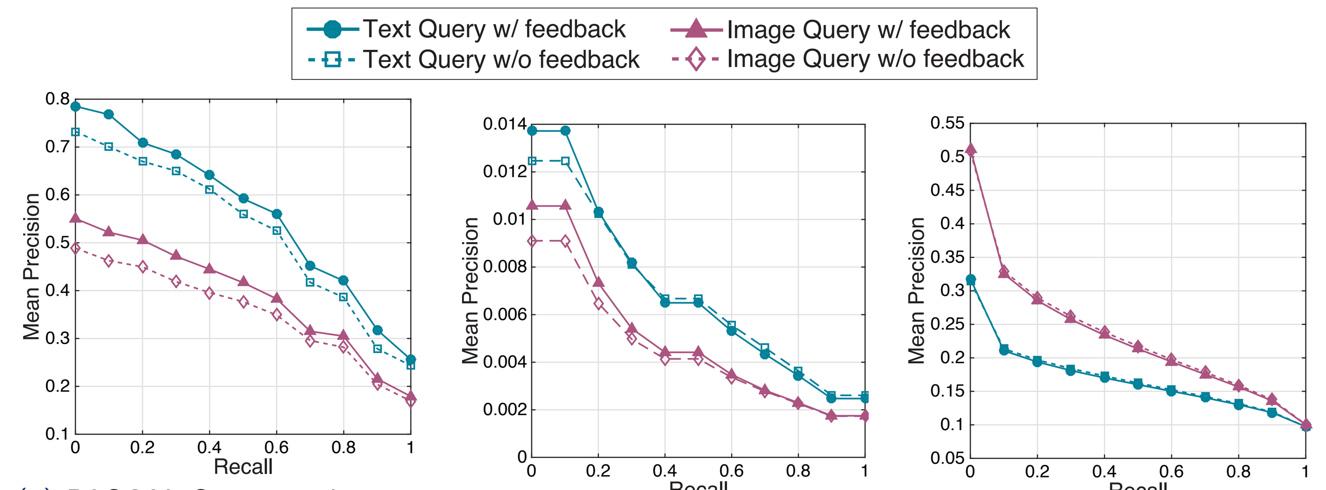


(a) PASCAL-Sentence dataset ($K = 27$)
(b) SUN dataset – Class label ground truth ($K = 58$)
(c) SUN dataset – Euclidean ground truth ($K = 58$)

Retrieval Experiments – Qualitative Analysis



Feedback Experiment – Quantitative Analysis



(a) PASCAL-Sentence dataset ($K = 27$)
(b) SUN dataset – Class label ground truth ($K = 58$)
(c) SUN dataset – Euclidean ground truth ($K = 58$)

Conclusions

- A Bayesian nonparametric framework for integrating multimodal data in a latent space.
- A retrieval system that can respond to cross-modal queries and an MCMC algorithm for inference.
- Formulation of user preferences as pseudo-images to alter the distribution of images in the latent space.

References

[1] Z. Ghahramani and T. L. Griffiths. Infinite latent feature models and the indian buffet process. In *NIPS*, pages 475–482, 2005.

[2] A. Farhadi, M. Hejrati, M. A. Sadeghi, P. Young, C. Rashtchian, J. Hockenmaier, and D. Forsyth. Every picture tells a story: Generating sentences from images. In *ECCV*, pages 15–29, 2010.

[3] G. Patterson and J. Hays. Sun attribute database: Discovering, annotating, and recognizing scene attributes. In *CVPR*, pages 2751–2758, June 2012.