Multimodal Learning in Loosely-organized Web Images

Kun Duan[1], David Crandall[1], and Dhruv Batra[2]

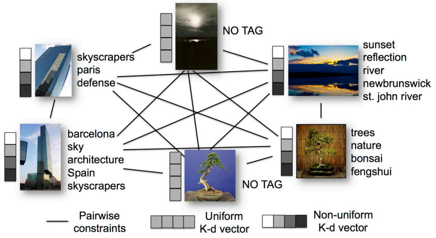
[1]Indiana University Bloomington, [2]Virginia Tech

INDIANA UNIVERSITY

VirginiaTech

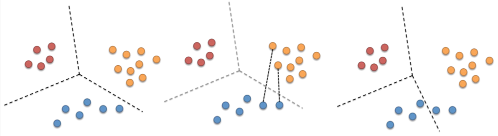
1. Overview

- **Motivation:** Photo-sharing websites have huge collections of images with (noisy, sparse) metadata like text tags, captions, timestamps and GPS. How can we organize these collections automatically?
- **Objective**: Cluster images using vision and noisy multimodal metadata.
- **Contributions:**
  1. General framework for loosely-supervised clustering for multimodal data with missing and incomparable features, using latent CRFs.
  2. Learn CRF parameters through metric learning and structured SVMs.
  3. Evaluate on large-scale online image datasets.



2. Multimodal Latent CRF Framework

- Generalize K-means by adding pairwise multimodal constraints:



Standard K-means… *…plus pairwise constraints…* *…yields constrained K-means.*

- Given instance features **X** with **m** different modalities, we solve for cluster centroids μ and cluster labels **Y**:

1. E-step: Fix μ, solve for **Y** jointly. Define a latent CRF to incorporate multimodal features in a single clustering framework:

$$\min_{\mu,\mathbf{y}} E(\{y_i\}|\{x_i\})$$

where:

$$E(\{y_i\}|\{x_i\}) = \sum_{i=1}^{N}\sum_{k=1}^{K} \mathbb{1}(y_i=k)\cdot\alpha(x_i^1,\mu_k) + \sum_{m=2}^{M}\sum_{i=1}^{N}\sum_{j=1}^{N}\beta_m(x_i^m,x_j^m)\cdot\mathbb{1}(y_i\neq y_j)$$

Distance function on primary feature channel

Similarity function on m^{th} feature channel

2. M-step: Fix **Y**, solve for μ using maximum likelihood estimation.

3. Parameter Learning

- **Learning similarity metrics:** We learn a similarity function for each channel using pairwise supervision, by applying ITML[1] and encoding distance metrics as diagonal Mahalanobis matrices.
- **Learning coefficients for similarity terms:** We learn similarity terms with a small held-out dataset with ground truth labels.
  - Formulate as a structured SVM learning problem:

Coefficients for similarity terms in latent CRF

Slack variable

$$\min_{\lambda,\mathbf{w},b} \lambda\|\mathbf{w}\|^2 + \xi,$$

such that,

$$E(\{\tilde{y}_i\}|\{x_i\}) - E(\{y_i\}|\{x_i\}) \geq \Delta(\{\tilde{y}_i\},\{y_i\}) - \xi,$$

$$\forall\{\tilde{y}_i\}\neq\{y_i\},\ \mathbf{w}\geq 0,\ \xi\geq 0.$$

Loss function

  - Our loss function is the *number of incorrect pairs* (Rand Index), which permits efficient loss-augmented inference.

$$\Delta(\{\tilde{y}_i\},\{y_i\}) = \sum_{i=1}^{N}\sum_{j=1}^{N}\mathbb{1}_{\tilde{y}_i=\tilde{y}_j\wedge y_i\neq y_j\vee\tilde{y}_i\neq\tilde{y}_j\wedge y_i=y_j}$$

4. Experimental Results

- **Datasets:** 3 labeled Flickr datasets (Landmarks, Groups, Sport, 10k images each); 1 unlabeled dataset (Activity, 30k images)
- **Visual features:** Bag-of-words SIFT histograms; **Metadata features:** binary tag occurrence vectors, GPS coordinates
- Tested with three different types of supervision, assuming differing types of training data:

Weak supervision

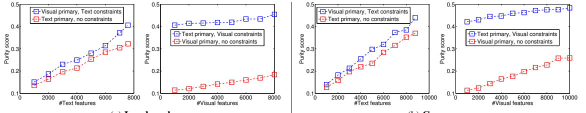
Use held-out set of pairs of images with same/different cluster annotations.

Purity: (i.e., proportion of the most common ground truth category in each cluster)

	Visual features	Text features	Visual+Text	Proposed (V+T)	Proposed (V+T+G)
Landmarks	0.1677 ± 0.0134	0.3224 ± 0.0335	0.3449 ± 0.0383	**0.4060** ± 0.0279	—
Groups	0.2508 ± 0.0097	0.3696 ± 0.0263	0.3955 ± 0.0341	0.4395 ± 0.0389	**0.4450** ±0.0353
Sport	0.1483 ± 0.0101	0.3454 ± 0.0386	0.3524 ± 0.0387	0.3713 ± 0.0309	**0.3965** ±0.0182

Inverse purity: (i.e., proportion of the most common cluster in each ground truth category)

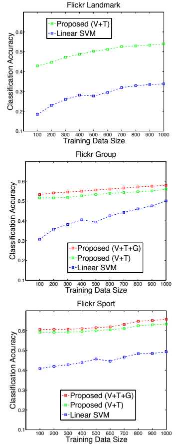
Landmarks	0.3163 ± 0.0180	0.4907 ± 0.0344	0.5297 ± 0.0227	**0.5611** ± 0.0210	—
Groups	0.4066 ± 0.0448	0.5893 ± 0.0275	0.5971 ± 0.0310	0.6010 ± 0.0322	**0.6336** ±0.0152
Sport	0.3707 ± 0.0411	0.6593 ± 0.0244	0.6789 ± 0.0175	0.6931 ± 0.0173	**0.7062** ±0.0190



(a) **Landmarks** (b) **Groups**

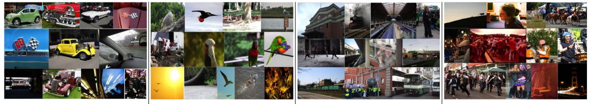
Loose supervision

Use ground-truth cluster labels for a subset of images.



Weak supervision on another dataset

Use parameters trained on one dataset (Sport) to cluster another dataset (Activity).



car show, California, people | bird, zoo, people, nature, wildlife | train, station, railway, people, show | people, festival, england, london

5. Summary and Conclusions

- Multimodal image clustering with visual features and sparse, noisy metadata, using latent CRFs.
- Learn feature distance functions and CRF parameters with varying degrees of supervision.

References:

[1] J. Davis, B. Kulis, P. Jain, S. Sra, and I. Dhillon. Information-theoretic metric learning. In *ICML*, 2007.

Acknowledgements: This work was supported in part by the National Science Foundation (CAREER IIS-1253549) and by the Indiana University Office of the Vice Provost for Research through the Faculty Research Support Program.