Multimodal Learning in Loosely-organized Web Images

Kun Duan¹, David Crandall², and Dhruv Batra³
¹Indiana University Bloomington, ²Virginia Tech

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:
     \[ \min_{\mu \in X, \mathcal{Y}} \sum_{i=1}^{N} \| y_i \|_{K_{\mu}} \]
     where:
     \[ E(\mathcal{Y}(\mu)) = \sum_{i=1}^{N} \sum_{j=1}^{M} \mathcal{Z}(y_i, y_j) = \sum_{i=1}^{N} \sum_{j=1}^{M} \mathcal{Z}(y_i, y_j) \]
     Distance function on primary feature channel
     \[ \mathcal{Z}(y_i, y_j) = \sum_{k=1}^{K} \mathcal{L}(y_i, y_j) \]
     Similarity function on \( m \)th feature channel
     \[ \mathcal{L}(y_i, y_j) = \sum_{k=1}^{K} \mathcal{Z}(y_i, y_j) \]
     \[ \mathcal{Z}(y_i, y_j) = \sum_{k=1}^{K} \mathcal{L}(y_i, y_j) \]
     \[ \mathcal{L}(y_i, y_j) = \sum_{k=1}^{K} \mathcal{Z}(y_i, y_j) \]

3. Parameter Learning
   • Learning similarity metrics: We learn a similarity function for each channel using pairwise supervision, by applying ITML^2 and encoding distance metrics as diagonal Mahalanobis metrics.
   • 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:
       \[ \min_{\lambda, \mathcal{A}, \mathcal{B}} \lambda \| \mathcal{A} \|_{1} + \xi \]
       slack variable
       \[ \text{such that,} \]
       \[ E(\mathcal{Y}(\mu)) - E(\mathcal{Y}(\mu)) - \Delta(\mathcal{Y}(\mu)) \geq \xi \]
       \[ \forall (y_i, y_j), \mathcal{A} \geq 0, \xi \geq 0, \]
       • Our loss function is the number of incorrect pairs (Rand Index), which permits efficient loss-approximated inference.
       \[ \Delta(\mathcal{Y}(\mu)) = \sum_{i=1}^{N} \sum_{j=1}^{M} \mathcal{Z}(y_i, y_j) \]

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.

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:

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.