Multimodal Learning in Loosely-organized Web Images

Kun Duan
Indiana University
Bloomington, Indiana
kduan@indiana.edu

David J. Crandall
Indiana University
Bloomington, Indiana
djcran@indiana.edu

Dhruv Batra
Virginia Tech
Blacksburg, Virginia
dbatra@vt.edu

Abstract

Photo-sharing websites have become very popular in the last few years, leading to huge collections of online images. In addition to image data, these websites collect a variety of multimodal metadata about photos including text tags, captions, GPS coordinates, camera metadata, user profiles, etc. However, this metadata is not well constrained and is often noisy, sparse, or missing altogether. In this paper, we propose a framework to model these "loosely organized" multimodal datasets, and show how to perform loosely-supervised learning using a novel latent Conditional Random Field framework. We learn parameters of the LCRF automatically from a small set of validation data, using Information Theoretic Metric Learning (ITML) to learn distance functions and a structural SVM formulation to learn the potential functions. We apply our framework on four datasets of images from Flickr, evaluating both qualitatively and quantitatively against several baselines.

1. Introduction

Online photo-sharing has become very popular in the last few years, generating huge collections of images on sites like Flickr, Picasa, and Instagram. As these datasets grow ever larger, a key challenge is how to organize them to allow for efficient navigation and browsing. For instance, we may want to discover the structure of photo collections by clustering images into coherent groups with similar objects, scenes, events, etc. in an automatic or semi-automatic way.

While image clustering has been studied extensively (*e.g.* [3, 21, 23] among many others), photo collections on modern photo-sharing sites introduce new opportunities and challenges. In addition to the images themselves, photos on these sites often include rich metadata that provide additional cues to the semantic content of the images, including text tags, timestamps, camera EXIF data, GPS coordinates, captions, and comments from other users. This metadata allows us to find connections between photos that are not obviously similar: a photo of the crowd at a candidate's political rally is clearly related to a photo of his or her campaign logo, but these photos exhibit almost no visual similarity. In such cases, similarities in the non-visual metadata may help: image tags and captions often contain useful keywords related to the content, activities, and context of the scene, while GPS coordinates and timestamps can be used to find photos taken nearby in space and time.

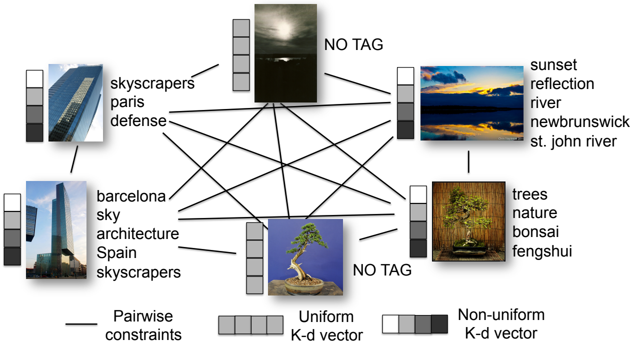


Figure 1: Latent Conditional Random Field model for two feature types. The primary features here are text tags, which are encoded as unary potentials, while visual features are the constraints (encoded in the pairwise potentials). Missing text tags yield uniform unary potentials.

Of course, metadata alone is not enough: two random photos tagged canon d50 are probably not related, while photos tagged with identical GPS and timestamps may be unrelated if taken on different floors of a large building. Moreover, metadata is typically not well constrained, and thus often missing, incomplete, ambiguous, or erroneous. For instance, some photos include detailed text tags, while others are tagged with unhelpful or noisy labels or are not tagged at all; even the most fastidious of photographers cannot list *all* possible tags that are relevant to an image. GPS coordinates are only collected by select devices like smartphones and are often hidden due to privacy concerns, so geo-tags typically appear on a small subset of images.

Here we present an approach for clustering large datasets with multimodal visual and non-visual features, and apply it to social photo collections. We are particularly interested in the incomplete and noisy nature of non-visual metadata: can modality features that are very sparse be used in a principled way to improve clustering performance? To solve this novel problem, we propose a generalization of the K-means algorithm using latent CRFs. Our method can be used in a fully unsupervised setting, or can use labeled training data if available, in contrast to supervised methods like SVMs that require significantly more training data. Our method is designed for cases where obtaining large labeled datasets is not possible, but annotating a small amount of training data is feasible. For example, in a large scale photo collection with millions of images, if the categories of interest are known in advance, one can manually annotate a few hundred instances, and apply our approach using this loosely-supervised information for organizing the rest.

As in traditional clustering (like K-means), we wish to assign each instance to a cluster, but the cluster identities (*e.g*. centroids) are themselves unknown and must also be inferred. We pose this problem using a Latent Conditional Random Field, in which each node in the graph corresponds to an image, and our goal is to mark each node with a cluster label. We pick one type of feature to be the *primary feature* and use it to define the CRF's unary potentials, which are functions of the distances from an image's primary feature to each latent cluster center. The other feature channels are considered to be *constraints* and appear as pairwise potentials in the CRF. These constraints tie together images with similar secondary features, encouraging them to be assigned to the same cluster. Incomplete, noisy, and heterogeneous features can thus be naturally incorporated into this model through these soft constraints. To perform clustering, we alternately solve for cluster assignments and cluster centers in a manner similar to K-means and EM, except that the E-step is much more involved, requiring inference on a CRF.

A challenge in clustering with noisy, multimodal features is how to define sensible distance metrics for the heterogeneous feature types, and how to weight them relative to one another. We address this problem by learning the distance and potential functions on a small amount of labeled training data we obtain from each category. In particular, we use Information Theoretic Metric Learning (ITML) [4] to learn the parameters of the distance metrics for constraint features, and use structural SVMs with the same training exemplars to learn the potential functions of the CRF. Our approach can still work for unsupervised cases, when obtaining labeled images is not feasible or no prior knowledge about the categories of interest is known; in this case, we can use a standard metric like L_2 distance, or a distance function learned on a different but similar dataset.

Finally, we evaluate our approach on three datasets from Flickr, with labeled ground truth and different types of features including visual, text, and GPS tags, and compare against baseline methods. We also test on a large unlabeled dataset, showing that our technique can find coherent events and activities in a completely unsupervised manner.

To summarize, the contributions of this paper are: (1) to propose a general framework for loosely-supervised clustering for multimodal data with missing features; (2) to apply metric learning and formulate a structural SVM problem for learning the structure of the latent CRF; and (3) to show that the approach can be used for unsupervised clustering on large-scale online image datasets.

2. Related Work

There is a vast literature on unsupervised and semi-supervised learning in the data mining community, and these techniques have been applied to organizing photos in a variety of contexts [5,6,11,12,21–23]. Two research threads are most closely related to this paper: multimodal modeling in image collections, and constrained clustering.

Multimodal modeling. McAuley and Leskovec [13] use relational image metadata (social connections between photographers) to model pairwise relations between images, and they apply a structural learning framework to solve the resulting labeling problem. While similar to our work in spirit, their formulation does not allow for missing metadata, and does not incorporate multimodal features (and does not use visual features at all). Rohrbach *et al* [15] propose a framework to recognize human activities in videos using both visual and detailed textual descriptions. Guillaumin *et al* [7] use a semi-supervised classifier on visual and text features for image classification; they allow missing class labels on training images, but do not allow for sparse features (they assume that all training images have text tags). In contrast, our model allows missing features in any modality channel, and learns the concepts in a *loosely supervised* manner (using just a small labeled training dataset to learn the parameters of our CRF).

Bekkerman and Jeon [3] perform unsupervised multimodal learning for visual features and text, but similarly do not attempt to handle sparse or missing features. Perhaps most relevant to our work is that of Srivastava and Salakhutdinov [16], who propose a multimodal learning method using deep belief networks. Their work allows for missing modalities on training instances by a sampling approach, but their technique can be expensive because it requires many different layers and also a lot of parameters. On the other hand, we propose a lightweight unsupervised learning framework which discovers clusters automatically, but that can still be used to build discriminative classifiers to predict missing modalities on new unseen images.

Constrained clustering. Several papers incorporate constraints into classical clustering algorithms like K-means.

Our approach can be thought of as constrained clustering, similar to HMRF-Kmeans [2] and related work [11, 12, 18], but there are key differences in motivation and formulation. We explicitly deal with missing features (which are quite common in web images) while these existing methods do not consider this problem. Intuitively, our framework only performs K-means updates (the "M-step") for one feature channel; when this type of feature is missing on some instances, K-means updates are calculated based on a subset of the network. Our work is related to Wagstaff *et al* [19] and Basu *et al* [2] who add "hard" constraints to the standard K-means framework, including "must-link" and "cannot-link" constraints between data points. In our application, where metadata is noisy and often inaccurate or ambiguous, such hard constraints are too strong; we instead use "soft" constraints that encourage instances to link together without introducing rigid requirements. Our models also allow different feature types in the pairwise constraints (e.g. some constraints may be defined in terms of tag relations, while others are defined using GPS, etc).

3. Loosely Supervised Multimodal Learning

We now present our approach for loosely supervised clustering in datasets with multimodal, sparse features. We assume that there are multiple feature types that are not comparable with one another, and observed values for some of these features on each instance in our dataset. For example, for online photos we may have visual features, text tags, and geotags, for a total of three feature modalities, and visual features are observable in all images but the others are available on just a subset. Our goal is to jointly consider all of this sparse and heterogeneous evidence when clustering.

3.1. Constrained Clustering Framework

We can think of our approach as a generalization of the classic K-means clustering algorithm. In K-means, we are given a dataset of instances $X = \{x_1, ..., x_N\}$, where each instance is a point in a d-dimensional space, $x_i \in \mathcal{R}^d$. Our goal is to assign one of K labels to each instance, i.e. to choose $y_i \in [1, ..., K]$ for each instance x_i, and to estimate K cluster centers $\mu = \{\mu_1, ..., \mu_K\}$, so as to minimize an objective function measuring the total distance of points from assigned centroids,

$$\min_{\mu, \mathbf{y}} \sum_{i=1}^{N} \sum_{k=1}^{K} \mathbb{1}(y_i = k) \, \|x_i - \mu_k\|^2, \tag{1}$$

where $\mathbf{y} = (y_1, ..., y_N)$ and $\mathbb{1}(\cdot)$ is an indicator function that is 1 if the given condition is true and 0 otherwise. Note that this formulation implicitly assumes that each instance can be represented by a point in a d-dimensional space, and that Euclidean distances in this space are meaningful.

In our approach, we assume that we have M different types of features, only a subset of which are observable in any given instance. Our dataset thus consists of a set of N instances, $\mathbf{X} = \{x_1, ..., x_N\}$, where each $x_i = (x_i^1, ..., x_i^M)$, and a given x_i^m is either a feature vector or $\emptyset$ to indicate a missing value. We treat one of these as the *primary* feature (we discuss how to choose the primary feature below) and consider the others as soft constraints, which tie together instances having similar values. We assume without loss of generality that the primary features have index $m = 1$. Any of these feature types (including primary) may be missing on a given instance. An illustration of our approach is shown in Figure 2. Now we can generalize the K-means energy function in equation (1) as,

$$\min_{\mu, \mathbf{y}} E(\{y_i\}|\{x_i\}), \tag{2}$$

with

$$\begin{aligned} E(\{y_i\}|\{x_i\}) = &\sum_{i=1}^{N} \sum_{k=1}^{K} \mathbb{1}(y_i = k) \cdot \alpha(x_i^1, \mu_k) \\ &+ \sum_{m=2}^{M} \sum_{i=1}^{N} \sum_{j=1}^{N} \beta_m(x_i^m, x_j^m) \cdot \mathbb{1}(y_i \neq y_j), \end{aligned} \tag{3}$$

and where $\alpha(\cdot, \cdot)$ is a distance function that defaults to 0 if a primary feature is missing,

$$\alpha(x_i^1, \mu_k) = \mathbb{1}(x_i^1 \neq \emptyset) \cdot \|x_i^1 - \mu_k\|^2,$$

and $\beta_m(\cdot, \cdot)$ is a function that measures the similarity between the m-th (non-primary) feature of two instances (described below), or is 0 if one or both of the features are missing. Intuitively, the first summation of this objective function is identical to that of the objective function of K-means in equation (1), penalizing distance from the primary features to the cluster centroids. If a primary feature is missing in a given instance, it does not contribute to the objective function (since any assigned label has equal cost). In the special case that there is exactly one feature type and it is always observable, equation (3) is equivalent to simple K-means in equation (1). The non-primary features add soft constraints through the second set of summations in equation (3), penalizing pairs of instances from being assigned to different clusters if they have similar features.

The objective function in equation (3) is a Latent Conditional Random Field model. Each instance (image) is a node in the CRF, and the goal is to label each node with a cluster identifier. The primary features define unary potentials, which give a cost for assigning a given node to each centroid, or a uniform distribution if the primary feature is missing. As in K-means, the cluster centroids are latent variables that must be estimated from data. Edges connect together pairs of instances where non-primary feature are

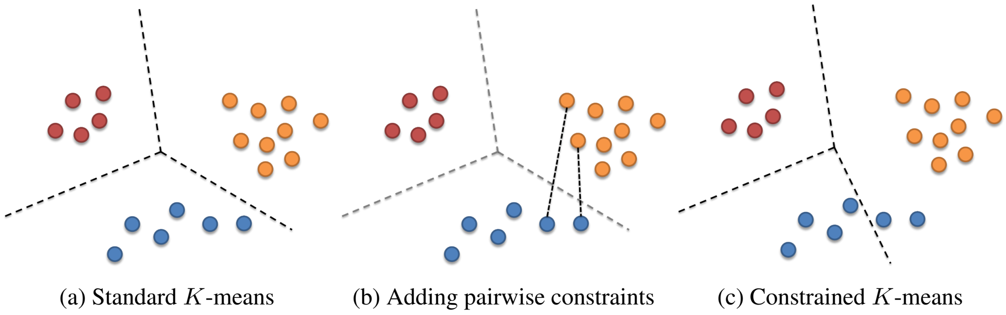


Figure 2: Illustration of our constrained clustering framework. (a) Standard K-means has only one feature type; (b) we add more feature types, which induce pairwise soft constraints between instances; (c) CRF inference balances evidence from all features in performing the clustering.

available, with pairwise potentials given by the β functions. To perform clustering in this framework, we must perform inference on the latent CRF. This is an optimization problem with two sets of unknown variables: the cluster centers μ and the cluster assignments $\mathbf{y}$. We use an EM-like coordinate descent algorithm to solve this problem, iteratively applying the following steps:

1. In the **E-step**, we fix μ and (approximately) solve for $\mathbf{y}$ by performing discrete optimization on the CRF using tree-reweighted message passing (TRW-S) [8].
2. In the **M-step**, we fix $\mathbf{y}$, and solve for each μ_k with simple maximum-likelihood estimation.

Note that these two steps are the familiar algorithm used in K-means, except that the E-step here involves jointly assigning cluster labels to the instances by performing inference on a CRF (instead of simply assigning each instance to the nearest cluster center as in K-means). The M-step is identical to that of K-means, except that here we ignore instances with missing primary features.

We can use this framework in different ways, depending on the amount of information available in a given application. In a *weakly supervised* setting, we assume that for some pairs of instances (in a held-out set), we know whether each pair belongs to the same class or a different class. We use these labels to learn the pairwise potentials as described in Section 3.2. We can learn a distance metric even when the constraint features are available but the primary feature is missing, or when the labeled set is in a different domain than the clustering application at hand. In a *loosely supervised* setting, we make the stronger assumption that a small subset of instances have ground-truth class labels, such that we can estimate the centroids using the small subset, and fix the centroid labels in that subset while solving for the rest.

3.2. Learning Pairwise Potentials

The clustering framework in Section 3.1 requires pairwise potential functions $\beta_m(\cdot,\cdot)$ to evaluate the similarity between two instances according to each feature type. These functions are critically important to clustering performance and thus we learn their parameters automatically. We define the pairwise potentials for each feature type m to have the following parametric form,

$$\beta_m(x_i^m, x_j^m)=\mathbb{1}(x_i^m\neq\emptyset\wedge x_j^m\neq\emptyset)\cdot(w_m\cdot d_m(x_i^m,x_j^m)+b), \tag{4}$$

where $d_m(\cdot,\cdot)$ is a (learned) distance function for the given feature type, w_m and b are scalar weight and bias terms, and the indicator function ensures $\beta_m(\cdot,\cdot)$ is clamped to 0 if either feature is missing. Learning the potential functions now involves estimating the distance function $d_m(\cdot,\cdot)$ for each feature type, and the weight and bias terms w_m and b; we estimate these in two separate steps.

Learning the distance functions. We assume that the distance functions are Mahalanobis distances,

$$d_m(x_i^m, x_j^m) = (x_i^m - x_j^m)^T A_m (x_i^m - x_j^m),$$

and thus we need only to estimate the matrices A_m. To do this, we use Information Theoretic Metric Learning (ITML) [4] to learn these matrices from pairwise supervision on the small labeled training data. For increased robustness to noise, we used diagonal Mahalanobis matrices.

Learning the potential function parameters. We wish to jointly estimate the $M-1$ feature weight parameters $\mathbf{w}=(w_2,...,w_M)$ and the bias term b in equation (4). We formulate this as a standard margin-rescaled structural SVM learning problem [17]. Let y_i and $\tilde{y}_i$ be the ground truth and predicted label of x_i, $E(\{y_i\}|\{x_i\})$ be the energy when the labelings are $\{y_i\}$ (in equation (3)); we minimize,

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

such that,

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

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

We define our loss function using *number of incorrect pairs*,

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

in other words, for each pair of instances in the dataset, we count how many of them were incorrectly assigned to different clusters and how many were incorrectly assigned to the same cluster. This definition of loss is the *Rand Index* [14], a popular evaluation metric in the clustering literature. We chose to use this metric (as opposed to other popular metrics like purity) because it allows the loss function to decouple into independent optimizations over each data point. We can then perform *loss-augmented inference* using the TRW-S algorithm [8] at training time, allowing for efficient inference in the inner loop of structured SVM training.

4. Experiments

We demonstrate our clustering method on four datasets collected from Flickr, three of which have ground-truth to allow for quantitative evaluation. In the fourth dataset, we show how our technique can be used to discover structure in large collections of images for which no ground truth exists.

4.1. Applications and datasets

We use four datasets of images from Flickr collected using the public API. To test the robustness of our approach in different settings, each of these datasets targets a different application of unsupervised clustering, and uses different feature types and ground truth collected in varying ways.

Landmarks. Our first dataset contains images from the ten most-photographed landmarks on Flickr, using the dataset from [10]. That paper clusters geo-tags to find highly-photographed places and learns discriminative classifiers for each of these landmarks. Here we test if our method can separate the landmarks in a less supervised manner, which could be useful in organizing large tourist photo collections around travel destinations. In this dataset we use only image features and text tags; we do not use GPS features because they were used to define the ground truth classes. We hide the ground truth, apply our clustering framework on image and tag features, and then compare the clustered results with the ideal clustering induced by the class labels. This **Landmarks** dataset includes 8,814 images.

Groups. Sites like Flickr let users contribute their photos to groups about user-defined topics. These groups have rich and varied themes, and the ability to categorize photos into groups automatically could be useful to help users organize their photos. We collected 1,000 images from each of 10 Flickr groups related to the following topics: aquarium, boat, bonsai, cars, Christmas, fireworks, food, penguins, skyscrapers, and sunsets. (These are the topics shown in Fig. 1 of [20]; unfortunately those authors could not share their dataset, so we found Flickr groups corresponding to the same topics and gathered our own images). We use visual, text, and geo-tag features in this **Groups** dataset.

Activities. We are also interested in clustering images according to human activities like attending a game, going to a museum, taking a hike, etc. Since these activities correspond to higher level semantics than simple actions like walking, running, etc., they are difficult to classify using visual features alone. (For instance, a picture of cars could be "car racing" if the cars are moving or "car show" if they are stationary, but the difference in visual appearance is subtle.) We thus use our multimodal clustering algorithm to incorporate visual, textual, and GPS features into this organization process. We collected two activity-related datasets. **Sport** consists of 10,000 images related to sporting events, which we collected by crawling 10 types of Flickr groups (American football, baseball, basketball, hockey, horse racing, marathons, NASCAR, football (soccer), swimming, tennis). These group labels give ground truth for evaluation. **Activity** includes about 30,000 random images from Flickr, which we use to qualitatively test our approach's ability to discover activities in unlabeled data. Here we use a large number of clusters ($K = 1000$) so that we can find coherent clusters despite the large number of outlier images.

In collecting the above datasets, we were careful to prevent "leaks" between class labels and the features used for clustering. For example, we did not use text features to define class labels, instead relying on geo-tags and group assignments. We also prevented any single photographer from dominating the datasets by sampling at most 5 photos from any single user. In general, about 80% of images have at least one text tag and about 10% of images have a geo-tag.

4.2. Features

On **Landmarks**, **Groups**, and **Sport**, we represent each image using histograms of visual words (using SIFT descriptors and a visual vocabulary of 500 words built using K-means). For the text features, we apply PCA on the binary tag occurrence vectors to reduce the dimensionality to 200. We learn a Mahalanobis distance for the text features using the method in Section 3.2 on the lower-dimensional space. For geo-tags, we use chord lengths on the sphere as the distance between two GPS coordinates. On the **Activity** dataset, we compute high-level features using object bank [9], and use image captions as the text features. Stop words are removed, the remaining words are stemmed, and we represent the text using binary occurrence vectors and again apply PCA to reduce the dimensionality to 200.

4.3. Results

As mentioned in Section 3.1, our framework can be applied in different ways depending on the type of ground truth available. We first evaluate under weak supervision,

Purity:

	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:

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

Table 1: Purity (top) and Inverse Purity (bottom) on three datasets with $K = 10$ clusters. Means and standard deviations are over 5 trials. (GPS information is not available for **Landmarks**.) Our multimodal approach significantly outperforms single modality baselines and combined feature baselines, both in terms of purity and inverse purity.

which assumes that we have pairs of exemplars which we know belong to either the same or different classes, and we use these to learn the pairwise distances and potential functions. We also evaluate under loose supervision, which makes the stronger assumption that we have some exemplars with ground-truth class labels, so that the primary feature centroids can also be initialized.

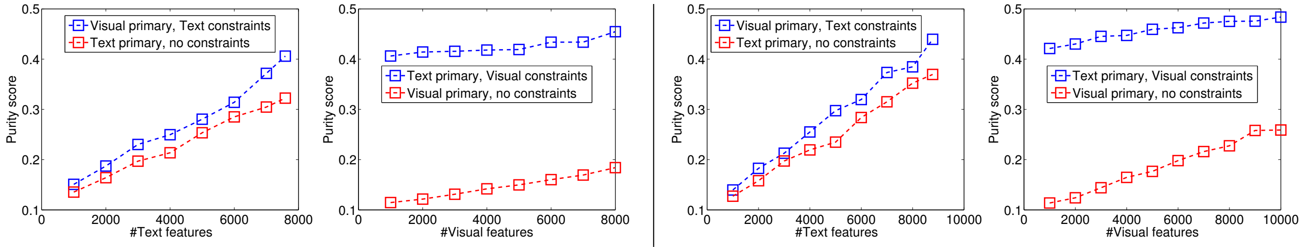
Weak supervision. Table 1 presents quantitative results for three datasets under weak supervision, using *purity* and *inverse purity* [1] as the evaluation metrics. For example, to compute purity, we calculate the percentage of instances within each estimated cluster that agree with the majority ground truth label of those instances. These numbers are averaged across all clusters to compute a final purity score. The table compares our method against several baselines: *Visual features* runs K-means on visual features only, *Text features* performs K-means using text features only, *Visual+Text* concatenates both features and performs K-means. Photos without tags are assigned random tags. *Proposed (V+T)* uses our approach with visual and text features, and *Proposed (V+T+G)* uses our approach with visual, text and GPS features. In each case we run 5 trials and report means and standard deviations, since results are non-deterministic due to the random initialization.

As shown in the table, our proposed method to incorporate (weak, sparse, noisy) multimodal data outperforms the baselines significantly. Visual features alone work relatively poorly (*e.g*. purity of about 0.17 for **Landmarks**), while text features are much more informative (0.32). Combining text and visual features together by simply appending the feature vectors and running K-means improves results slightly (0.34), while combining visual and text features in our framework significantly outperforms all of these baselines (0.41). Much of this improvement may come from our technique's ability to better handle photos that do not have text tags (about 20% of photos): when we exclude photos having no tags, the text-only K-means baseline increases to 0.3705 for **Landmarks** and 0.4567 for **Groups**. Finally, adding GPS features results in a modest additional gain.

We use text as the primary feature in the above experiments. We have found that the choice of primary feature is important, due to the different roles that the unary and pairwise potentials play in the constrained clustering framework. Intuitively, the pairwise constraints only depend on whether the labelings of two neighbors are the same, while the unary potentials encourage each node to explicitly select one of the K labels. It is thus easier for a labeling of the nodes to minimize the pairwise cost than the unary cost. To understand this better, we tested each of the two feature types (visual and text) in isolation as unary or pairwise constraints. Results of using only a unary term were already presented above, in the first two columns of Table 1; we tested the pairwise potentials in isolation by fixing the unary potentials to be uninformative uniform distributions. On **Landmarks**, switching visual features from primary to pairwise features causes purity to change from 0.1677 to 0.1462, a drop of 13%, while switching text features from primary to pairwise drops the purity by 31% from 0.3224 to 0.2223. This result suggests that we should select the "strongest," most informative feature as the primary.

Figure 3 studies how sparsity of primary and secondary and text and visual features affects results, by hiding features of varying numbers of images. For each dataset, the left plot compares results of using a subset of text features as the primary and no constraint features (red), with using all visual features as primary and subsets of text features as constraints (blue). The red line is thus the same as simple K-means, where images without text features are randomly assigned to a cluster. The right plot shows a similar comparison but with the roles of the text and visual features swapped. We see of course that more observations lead to better performance, with best results when using all available text as primary features and all visual features as constraints. But the results also highlight the flexibility of our approach, showing that multi-modal features (blue lines) significantly improve performance over a single feature type (red lines), even when only a small percentage of photos have the feature.

Loose supervision. We used small labeled subsets of different sizes to evaluate the loosely supervised paradigm,



(a) **Landmarks**

(b) **Groups**

Figure 3: Clustering performance as a function of number of images with different types of features. Red lines use primary features for only a subset of images and do not use constraints (*i.e.* as in classic K-means). Blue lines use our multimodal clustering framework, incorporating primary features for all images and a subset of images with constraint features. For each dataset, purity in the left plot is calculated using all images, while in the right plot it is calculated using images having tags.

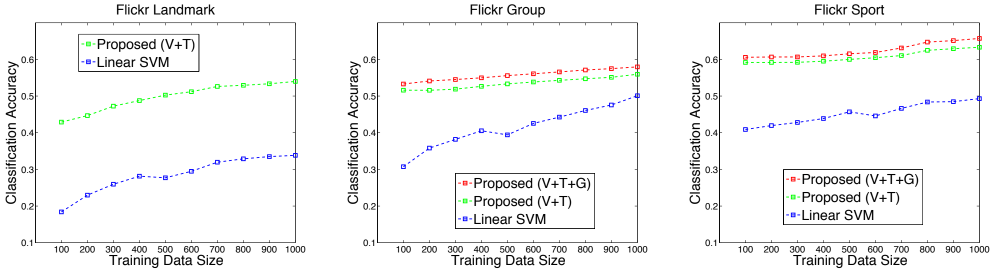


Figure 4: Classification performance comparisons with loose supervision on training sets of increasing sizes, using **Landmarks** (left), **Groups** (middle), and **Sport** (right). Linear SVM baseline is trained on concatenated visual and text features.

and evaluate using classification accuracy. We used linear SVMs trained on visual and text features as baseline methods, with the classifier parameters chosen according to 5-fold cross validation on the training data. Figure 4 shows that our proposed loosely supervised method outperforms SVM classifiers given the same amount of supervision, especially when the available training data is only a small percentage of the entire dataset. For instance, on **Landmarks**, our technique can achieve about 54% classification accuracy (relative to 10% random baseline) with 1,000 labeled exemplars, versus just 33% for a trained SVM using the same features and training set.

Qualitative results. Figure 5 presents sample clustering results for the **Landmarks**, where in each group we show the images closest to the cluster centroid and the most frequent tags in the cluster. Figure 6 presents sample clusters from our **Activity** dataset of 30,000 images, showing that the algorithm has discovered intuitively meaningful activity and event clusters like car shows, wildlife, festivals, beaches, etc. Since we do not have labeled ground truth for this dataset, we simply used the learned parameters from **Sport**.

5. Summary and Conclusion

We proposed a multimodal image clustering framework that incorporates both visual features and sparse, noisy metadata typical of web images. Our approach is loosely supervised, and is reminiscent of the standard K-means algorithm: one feature is used as the primary feature in K-means-style updates, while other features are incorporated as pairwise constraints. The proposed approach is flexible and can be applied under different degrees of supervision, including when no training data is available at all, and when features are missing. In future work, we plan to incorporate other types of constraints in the graphical model, and to apply our approach to various applications (*e.g.* automatic image annotation and recommendation).

Acknowledgments. This work was supported in part by a National Science Foundation CAREER award (IIS-1253549) and by the IU Office of the Vice Provost for Research through the Faculty Research Support Program.

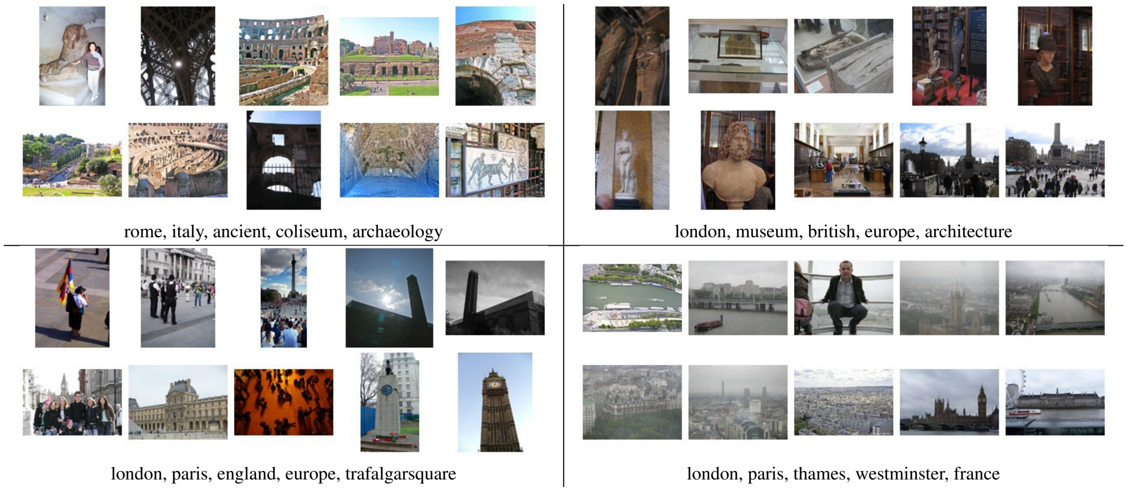
References

[1] E. Amigo, J. Gonzalo, J. Artiles, and F. Verdejo. A comparison of extrinsic clustering evaluation metrics based on formal constraints. *Inform. Retrieval*, 12(4), 2009. 6

[2] S. Basu, M. Bilenko, and R. Mooney. A probabilistic framework for semi-supervised clustering. In *KDD*, 2004. 3

[3] R. Bekkerman and J. Jeon. Multi-modal clustering for multimedia collections. In *CVPR*, 2007. 1, 2

[4] J. Davis, B. Kulis, P. Jain, S. Sra, and I. Dhillon. Information-theoretic metric learning. In *ICML*, 2007. 2, 4



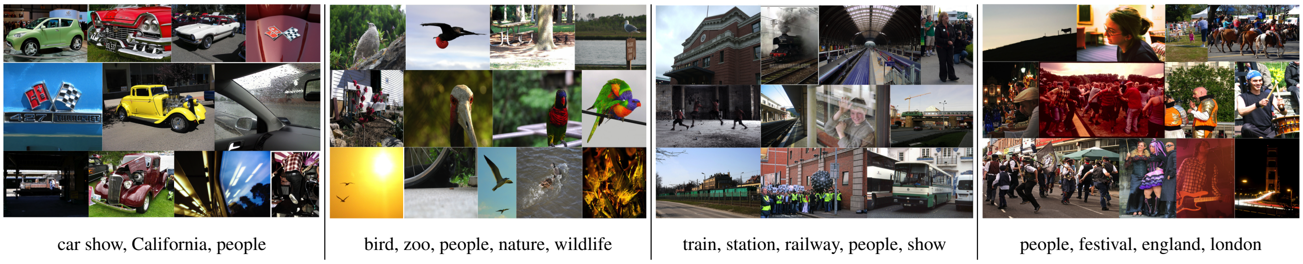
rome, italy, ancient, coliseum, archaeology

london, museum, british, europe, architecture

london, paris, england, europe, trafalgarsquare

london, paris, thames, westminster, france

Figure 5: Sample landmark clusters discovered automatically by our algorithm.



car show, California, people

bird, zoo, people, nature, wildlife

train, station, railway, people, show

people, festival, england, london

Figure 6: Some activities discovered by our algorithm.

[5] B. Gao, T.-Y. Liu, T. Qin, X. Zheng, Q. Cheng, and W.-Y. Ma. Web image clustering by consistent utilization of visual features and surrounding texts. In *ACM Multimedia*, 2005. 2

[6] J. Goldberger, S. Gordon, and H. Greenspan. Unsupervised image-set clustering using an information theoretic framework. *IEEE Trans. Image Proc.*, pages 449–458, 2006. 2

[7] M. Guillaumin, J. Verbeek, and C. Schmid. Multimodal semi-supervised learning for image classification. In *CVPR*, 2010. 2

[8] V. Kolmogorov. Convergent tree-reweighted message passing for energy minimization. *PAMI*, 28(10), 2006. 4, 5

[9] L.-J. Li, H. Su, E. P. Xing, and L. Fei-Fei. Object bank: A high-level image representation for scene classification & semantic feature sparsification. In *NIPS*, 2010. 5

[10] Y. Li, D. Crandall, and D. Huttenlocher. Landmark classification in large-scale image collections. In *ICCV*, 2009. 5

[11] Z. Li, J. Liu, and X. Tang. Constrained clustering via spectral regularization. In *CVPR*, 2009. 2, 3

[12] Z. Lu and M. Carreira-Perpiñán. Constrained spectral clustering through affinity propagation. In *CVPR*, 2008. 2, 3

[13] J. J. McAuley and J. Leskovec. Image labeling on a network: Using social-network metadata for image classification. In *ECCV*, 2012. 2

[14] W. Rand. Objective criteria for the evaluation of clustering methods. *J. Am. Statist. Assoc.*, 66(336):846–850, 1971. 5

[15] M. Rohrbach, M. Regneri, M. Andriluka, S. Amin, M. Pinkal, and B. Schiele. Script data for attribute-based recognition of composite activities. In *ECCV*, 2012. 2

[16] N. Srivastava and R. Salakhutdinov. Multimodal learning with deep Boltzmann machines. In *NIPS*, 2012. 2

[17] I. Tsochantaridis, T. Joachims, T. Hofmann, and Y. Altun. Large margin methods for structured and interdependent output variables. *JMLR*, 6:1453–1484, 2005. 4

[18] K. Wagstaff and C. Cardie. Clustering with instance-level constraints. In *ICML*, 2000. 3

[19] K. Wagstaff, C. Cardie, S. Rogers, and S. Schrödl. Constrained k-means clustering with background knowledge. In *ICML*, 2001. 3

[20] G. Wang, D. Hoiem, and D. A. Forsyth. Learning image similarity from Flickr groups using fast kernel machines. *PAMI*, pages 2177–2188, 2012. 5

[21] Y. Yang, D. Xu, F. Nie, S. Yan, and Y. Zhuang. Image clustering using local discriminant models and global integration. *IEEE Trans. Image Proc.*, pages 2761–2773, 2010. 1, 2

[22] J. Yu, M. Wang, and D. Tao. Semisupervised multiview distance metric learning for cartoon synthesis. *IEEE Trans. Image Process.*, pages 4636–4648, 2012. 2

[23] X. Zheng, D. Cai, X. He, W. Ma, and X. Lin. Locality preserving clustering for image database. In *MM*, 2004. 1, 2