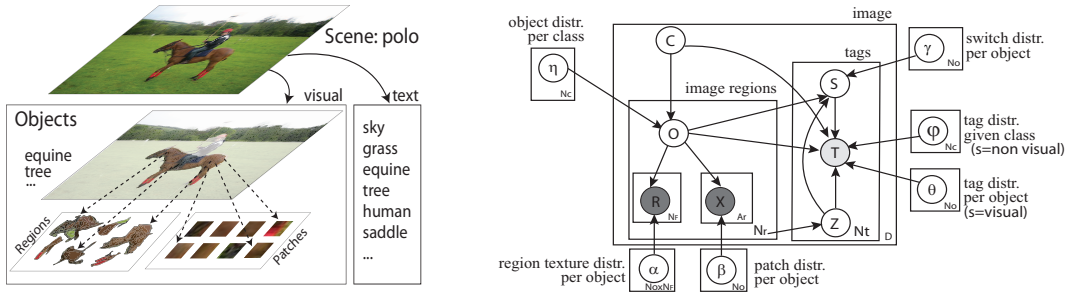
Towards coherent modeling of scene segmentation, annotation and classification

Li-Jia Li & Li Fei-Fei*, Princeton University, {jial, feifeili}@cs.princeton.edu

Given a real-world image such as the one in Fig.1(a), humans can easily extract a large amount of visual information with sometimes exquisite details in a matter of a fleeting moment. Computers today, on the other hand, are far from performing at this level. In this work, we focus on three important recognition tasks for complex real-world scenes: segmentation, annotation, and classification. *Segmentation* is defined as the task of delineating images based on meaningful components, namely visible objects such as trees, human, etc. *Annotation* involves two types. For each segmented region in the image, we annotate it with an object or concept label. In addition, annotation can also involve generating labels for an image without a corresponding image region, such as wind, a highly relevant concept for a sailing event picture. Finally, *classification* is the task of assigning a category label to the entire image that often denotes a higher-level concept such as a sport event, a natural scene class, etc. We present here a probabilistic model that achieves these three tasks in one coherent framework (Fig.1(a) is an example result).



(a) (b)

Figure 1: **a.** An example of what our model can understand given an unknown image. At the scene level, the image is classified as a 'polo' scene. A number of objects can be inferred and segmented by the visual information in the scene, hierarchically represented by object regions and feature patches. In addition, several tags can be inferred based on the scene class, including the visually unobserved object 'saddle'. **b.** A graphical model representation of our generative model. Nodes represent random variables and edges indicate dependencies. The variable at the right lower corner of each box denotes the number of replications. The box indexed by N_r represents the visual information of the image, whereas the one indexed by N_t represents the textual information (i.e. tags). N_c, N_o, N_x, $N_{f_i}, i \in 1, 2, 3, 4$ denote the numbers of different scenes, objects, patches and regions for region feature type i respectively. Hyperparameters of the distributions around the image block are omitted for clarity.

1 Contributions and related work

Total scene understanding. Most of the earlier object and scene recognition work offers a single label to an image, e.g. an image of a panda, a car or a beach. Some go further in offering a list of annotations without localizing where in the image each annotation belongs (e.g. [4]). A few concurrent segmentation and recognition approaches have suggested more detailed decomposition of an image into foreground object and background clutter. But all of them only apply to a single object or a single type of object (e.g. [2, 6]). Our proposed model captures the co-occurrences of objects and high-level scene classes. Recognition becomes more accurate when simultaneously recognizing different semantic components of an image, allowing each component to provide contextual constraints to facilitate the recognition of the others. *Our model can recognize and segment multiple objects as as well as classify scenes in one coherent framework.*

Flexible and unsupervised learning. Learning scalability is a critical issue when considering practical applications of computer vision algorithms. For learning a single, isolated object, it is feasible to obtain labeled data. But as one wishes to understand complex scenes and their detailed object components, it becomes increasingly labor-intensive and impractical to obtain labeled data. Fortunately the Internet offers a large resource of images and their tags. *We propose a framework for unsupervised learning from Internet images and tags (i.e. flickr.com), hence offering a scalable approach with no additional human labor.*

Robust representation of the noisy, real-world data. While the flickr images and tags provide a tremendous data resource, the caveat for exploiting such data is the large amount of noise in the user labels. The noisy nature of the labels is reflected in the highly uneven number and the quality of flickr tags: using a 'polo' image as an example, many

*presenting author

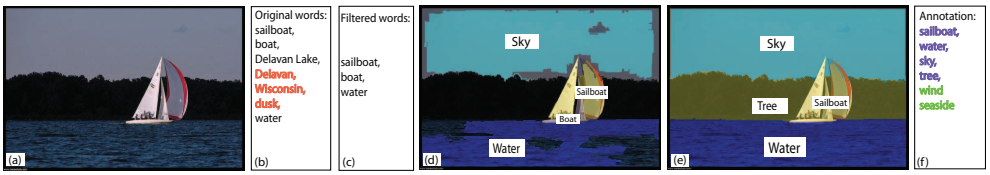


Figure 2: Walk-through of the learning process. **(a)** The original image. **(b)** The original tags from Flickr.com. Some of the irrelevant tags are colored in red. **(c)** Output of Step 1: Tags after the Wordnet pruning. While some of the irrelevant tags are pruned out, some relevant tags are missing, such as 'tree' for the image, and 'wind' for the sailing class. **(d)** Output of Step 2: Using an unsupervised clustering algorithm for each object or concept class (i.e. the filtered words in (c)) [2], we obtain an initial model for each object class. The training image is then partly annotated by using these models. Different object concepts are colored differently. The annotation results (i.e. tags) are overlayed on the corresponding regions. Note that there is a background class denoted by the black colored regions in the figure. This is the result of the highly conservative initialization strategy, allowing only very few regions in the training images to be annotated. The missing regions and tags will be recovered at a latter stage. **(e):** Output of Step 3: After training the hierarchical model by integrating all object concepts and user tags, the image is completely and more precisely segmented. **(f):** Final annotation proposed by our approach. Blue tags are predicted by the visual component ($S = visual$). Green tags are generated from the top down scene information learned by the model ($S = nonvisual$).

tags do not have obvious visual correspondences (e.g. 'pakistan', 'adventure'); some tags can be incorrect (e.g. 'snow', 'mountain'); and visually salient tags are often missing (e.g. 'grass', 'human'). *Our generative model offers, for the first time, a principled representation to account for noise related to either erroneous or missing correspondences between visual concepts and textual annotations.*

2 Approach

A coherent probabilistic model. Fig.1(b) denotes the graphical representation of our model. It describes the scene of an image through two major components: the visual component, where a scene consists of objects characterized by patches and region features; and a textual component where visually irrelevant information of the scene is captured. A switch variable (S) enables a principled joint modeling of images and texts and a coherent prediction of what tags are visually relevant. These two types of information are then tied together by a high-level class variable (C) that sends top-down information to influence the modeling of both visual and textual data.

An unsupervised learning framework. We derive a collapsed Gibbs sampling algorithm for learning the parameters of the model. During training, only the image patches, over-segmented regions and the noisy flickr user tags are observed. This would potentially introduce a problem of resolving the identity of objects that are always occurring together in a scene, such as water and sailboat for the sailing class. To prevent such a case, we introduce an automatic initialization step which provides a handful of labeled regions to seed the training process. Fig.2 illustrates the outcome of the different learning stages of an image during the training process.

3 Experiments and Results

We test our approach on 8 scene categories suggested in [5]: *badminton, bocce, croquet, polo, rock climbing, rowing, sailing, snowboarding*. We evaluate our algorithm on three different tasks: image level classification, image annotation, and pixel level segmentation. In the 8-way scene classification task, our model achieves an average of 54% accuracy (measured by the mean of the diagonal entries of the confusion table). This is in contrast with three other models tested using the same feature inputs: a baseline bag of words model [3], the unsupervised object recognition model used in our initialization stage [2] and a state-of-the-art image-texts model [1]. In the annotation task, we compare our algorithm with the state-of-the-art system presented in [4]. Precision-recall analysis shows a significant advantage of our model over [4]. Finally for the segmentation task, we compare our results with [2] and show superior quantitative results.

References

[1] D. Blei and M. Jordan. Modeling annotated data. *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, 2003.

[2] L. Cao and L. Fei-Fei. Spatially Coherent Latent Topic Model for Concurrent Segmentation and Classification of Objects and Scenes. *ICCV 2007*, 2007.

[3] L. Fei-Fei and P. Perona. A Bayesian hierarchy model for learning natural scene categories. *Computer Vision and Pattern Recognition*, 2005.

[4] J. Li and J. Wang. Automatic Linguistic Indexing of Pictures by a statistical modeling approach. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 25(9):1075–1088, 2003.

[5] L.-J. Li and L. Fei-Fei. What, where and who? classifying events by scene and object recognition. In *Proc. ICCV*, 2007.

[6] B. Russell, A. Efros, J. Sivic, W. Freeman, and A. Zisserman. Using multiple segmentations to discover objects and their extent in image collections. *Proc. CVPR*, 2006.