Semantic Context Learning with Large-Scale Weakly-Labeled Image Set

Yao Lu, Wei Zhang*, Ke Zhang, and Xiangyang Xue
School of Computer Science, Fudan University
Shanghai, China
{yaolu, weizh, k_zhang, xyxue}@fudan.edu.cn

ABSTRACT

There are a large number of images available on the web; meanwhile, only a subset of web images can be labeled by professionals because manual annotation is time-consuming and labor-intensive. Although we can now use the collaborative image tagging system, e.g., Flickr, to get a lot of tagged images provided by Internet users, these labels may be incorrect or incomplete. Furthermore, semantics richness requires more than one label to describe one image in real applications, and multiple labels usually interact with each other in semantic space. It is of significance to learn semantic context with large-scale weakly-labeled image set in the task of multi-label annotation. In this paper, we develop a novel method to learn semantic context and predict the labels of web images in a semi-supervised framework. To address the scalability issue, a small number of exemplar images are first obtained to cover the whole data cloud; then the label vector of each image is estimated as a local combination of the exemplar label vectors. Visual context, semantic context, and neighborhood consistency in both visual and semantic spaces are sufficiently leveraged in the proposed framework. Finally, the semantic context and the label confidence vectors for exemplar images are both learned in an iterative way. Experimental results on the real-world image dataset demonstrate the effectiveness of our method.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing

General Terms

Algorithms, Experimentation, Performance

Keywords

Image Annotation, Semantic Context, Large Scale, Weakly Labeled

*Corresponding Author: weizh@fudan.edu.cn

CIKM'12, October 29–November 2, 2012, Maui, HI, USA.

1. INTRODUCTION

Automatic image annotation is an efficient way for retrieving and managing numerous images on the web. In the task of image annotation, machine learning techniques are often used to learn classifiers from the labeled training images. Since annotating training samples by professionals is time-consuming and labor-intensive, only a subset of large-scale image dataset can be labeled manually. Although we can now use the collaborative image tagging system, e.g., Flickr, to get a lot of tagged images provided by Internet users, the tags of collaboratively-tagged images may not have exact semantics because the Internet users may tag the images according to their personal perceptions or social backgrounds.

In recent years, many algorithms on image annotation have been proposed. [13] introduced a technique for image annotation which used low-level visual features and a combination of basic distances called JEC to find the nearest neighbors of any given image; [7] proposed a method to annotate and retrieve images by learning one relevance model from a set of labeled samples; [9] proposed a weighted nearest-neighbor model based on neighbor rank or distance metric by maximizing the log-likelihood of the label predictions on training images. However, above methods did not allow for label-label correlation which is important to the performance in multi-label learning task. In real-world applications, semantics richness requires more than one label to sufficiently describe an image, and multiple labels usually interact with each other in semantic space. To capture the inter-label correlation for multi-label annotation application, [18] presented a measurement of the relationship between semantic concepts using the square root of Jensen-Shannon divergence between the corresponding visual language models; [4] proposed a hierarchical context model that captured object co-occurrences and spatial relationships among more than a hundred categories by a tree structure; [6] constructed a topic network to precisely characterize inter-topic (inter-label) contexts; [19] proposed a joint multi-label multi-instance learning model which captured the correlations between labels based on hidden conditional random fields. However, it is still unclear how to efficiently leverage unlabeled or weakly-labeled samples for multi-label learning in above methods.

To exploit weakly-labeled images available on the web for multi-label learning, [11] proposed a formulation to implement various tag analysis tasks in a unified framework, but the inter-label correlation was omitted in [11]. [10] introduced an image retagging method to improve the quality of the tags associated with social images in terms of content relevance; [15] proposed a bipartite graph reinforcement model for web image annotation, where a reinforcement algorithm was performed on the bipartite graph to re-rank the candidates; [17] defined the candidate annotations as the states of a Markov chain and formulated the annotation refinement process

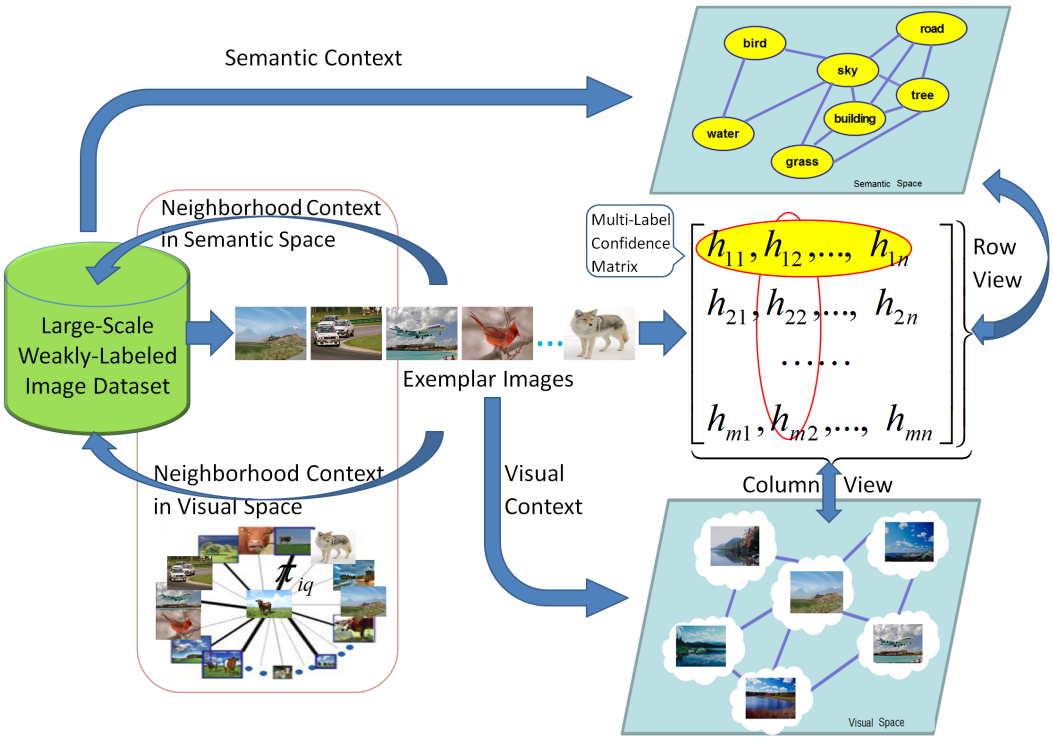

Figure 1: The framework of our model.

as a Markov process; [2] proposed a framework to improve the retrieval performance by refining noisy tags of a group of Flickr photos; [22] proposed a method to refine image labels by considering label correlation, content consistency, low-rank, and error sparsity. Above methods should be given a known semantic context as input, and how to learn semantic context from large-scale weakly-labeled image dataset is still not mentioned.

Semi-supervised learning is one way to learn from labeled and unlabeled samples [1]. In [23], a semi-supervised learning method was proposed to label data via a Gaussian random field model where the label of each datum was computed as the average of its neighbors; [8] introduced a semi-supervised learning scheme to propagate labels through images by constructing approximations to the eigenvectors of the graph Laplacian; [12] proposed a technique to make semi-supervised learning practical on large-scale dataset by seeking anchor points to construct a large adjacent graph. The frameworks in [1, 23, 8, 12] were not designed for multi-label learning, thus semantic context was not considered in above works. Recently, [16] proposed a sparse graph-based semi-supervised learning method to boost the performance of each concept detector in semantic space; [20] also proposed a graph-based learning framework in the setting of semi-supervised learning with multiple labels. Although [16, 20] allowed for multi-label learning, the semantic context should be provided as an input, instead of being learned from the image set automatically.

In this paper a novel method is developed to predict the labels for images by learning semantic context in a semi-supervised framework based on a large-scale web image dataset. By investigating the label confidence matrix for image exemplars from different perspectives, our method sufficiently leverages visual context, semantic context, and neighborhood consistency in both visual and semantic spaces. To address the scalability issue, a small number of exemplar images are first obtained to cover the whole data cloud, then the label vector of each image is estimated as a local combination of the label vectors of these exemplars. The semantic context and the label confidence vectors for exemplar images are both learned in an iterative way.

The rest of this paper is organized as follows: Section 2 gives the overview of semantic context learning and image annotation framework with large-scale weakly-labeled dataset. In Section 3, we formulate the proposed model. Experimental results on the real-world web image dataset are shown in Section 4. Finally, we conclude this paper in Section 5.

2. OVERVIEW OF OUR FRAMEWORK

Fig.1 gives the framework of our model. Firstly, a small number of exemplar images are selected to cover the whole data cloud, then the neighborhood contexts between samples and exemplars are preserved when mapping images from visual feature space to semantic label space. We investigate the label confidence matrix for image exemplars from column and row views, which should be consistent with visual context and semantic context, respectively. Visual context can be derived from the set of image exemplars and semantic context can be learned by our algorithm on large-scale weakly-labeled image dataset. We predict the label vectors for images by leveraging semantic context, visual context, and neighborhood context simultaneously.

3. THE PROPOSED MODEL

Let $\{(\mathbf{x}^1, \mathbf{y}^1), ..., (\mathbf{x}^l, \mathbf{y}^l)\}$ and $\{\mathbf{x}^{l+1}, ..., \mathbf{x}^{l+u}\}$ denote l labeled and u unlabeled images respectively, $\mathcal{C} = \{c_1, ..., c_m\}$ be the semantic lexicon of m concepts, and $\mathbf{y}^i = [\mathbf{y}^i_1, \ldots, \mathbf{y}^i_m]^\top \in \{0, 1\}^m$. If the concept c_s is associated with $\mathbf{x}^i$, then $\mathbf{y}^i_s = 1 (s = 1, \ldots, m)$; otherwise, $\mathbf{y}^i_s = 0$. Furthermore, let $\mathbf{h}^i \in [0, 1]^m$ denote the label confidence vector for the image $\mathbf{x}^i$, and the $s-$th

element of $\mathbf{h}^i$ measures the probability that the image $\mathbf{x}^i$ has the concept c_s. Our goal is to predict the label vectors for images by learning semantic context with the large-scale web image dataset.

We employ K-means clustering algorithm to seek a small number of exemplars covering the total data cloud. Suppose that n clusters are obtained, and the sample closest to the center for each cluster is regarded as the corresponding exemplar, then we get n exemplars $\{\dot{\mathbf{x}}_1, ..., \dot{\mathbf{x}}_n\}$. Like [14, 12, 21, 3], each image is approximately reconstructed as a local combination of the exemplars:

$$\mathbf{x}^i \approx \sum_{q \in \langle \mathbf{x}^i \rangle} \pi_{iq} \dot{\mathbf{x}}_q, \tag{1}$$

where $\langle \mathbf{x}^i \rangle$ is the index set of k nearest exemplars for $\mathbf{x}^i$, and π_{iq} can be estimated as follows:

$$\begin{aligned} \min \quad & \frac{1}{2}\big\|\mathbf{x}^i - \sum_{q \in \langle \mathbf{x}^i \rangle} \pi_{iq} \dot{\mathbf{x}}_q\big\|^2, \\ s.t. \quad & \pi_{iq} \geq 0, \sum_{q \in \langle \mathbf{x}^i \rangle} \pi_{iq} = 1 \end{aligned} \tag{2}$$

If we preserve the neighborhood contexts when images are mapped from visual space to semantic space, then the label confidence vector for each image $\mathbf{x}^i$ can be approximated as a local combination of the labels of exemplars as well:

$$\mathbf{h}^i = \dot{H}\vec{\alpha}^i, \tag{3}$$

where $\dot{H} = [\dot{\mathbf{h}}_1, ..., \dot{\mathbf{h}}_n] \in [0,1]^{m \times n}$, and $\dot{\mathbf{h}}_q \in [0,1]^m$ is the label confidence vector of the exemplar $\dot{\mathbf{x}}_q, (q = 1, ..., n)$; $\vec{\alpha}^i = [\pi_{i1}, ..., \pi_{in}]^\top$, and $\pi_{iq} = 0$ if $q \notin \langle \mathbf{x}^i \rangle$.

Thus the label confidence vector of all samples can be obtained as $H = [\mathbf{h}^1, ..., \mathbf{h}^{l+u}] = \dot{H}A$, where $A = [\vec{\alpha}^1, ..., \vec{\alpha}^{l+u}] \in R^{n \times (l+u)}$. For those labeled samples, the corresponding label confidence vector should be consistent with the given labels to some extent. Thus, we minimize the following loss function:

$$\begin{aligned} \min_{\dot{H}} \quad & \sum_{i=1}^{l} \big\|\mathbf{y}^i - \mathbf{h}^i\big\|^2 \\ & = Tr\big((Y - \dot{H}A_l)^\top (Y - \dot{H}A_l)\big) \end{aligned} \tag{4}$$

where $Y = [\mathbf{y}^1, ..., \mathbf{y}^l] \in \{0,1\}^{m \times l}$, $\dot{H} = [\dot{\mathbf{h}}_1, ..., \dot{\mathbf{h}}_n] \in [0,1]^{m \times n}$, and $A_l \in R^{n \times l}$ is the sub-matrix according to the labeled subset.

To capture the visual context, we define a similarity matrix $S \in R^{(l+u) \times (l+u)}$ measuring the visual similarities between image pairs. The matrix S is large-scale. Inspired by [12], we can approximate the weight matrix S as follows:

$$\widetilde{S} = A^\top \Lambda^{-1} A \tag{5}$$

where the diagonal matrix $\Lambda \in R^{n \times n}$ is defined as $\Lambda_{ii} = \sum_{q=1}^{l+u} A_{iq}$. The visual and semantic consistency can be achieved by:

$$\begin{aligned} \min_{\dot{H}} \quad & \frac{1}{2} \sum_{i,j} \widetilde{S}_{ij} \big\|col(H,i) - col(H,j)\big\|^2 \\ & = Tr\big(HL_sH^\top\big) = Tr\big(\dot{H}AL_sA^\top\dot{H}^\top\big) \end{aligned} \tag{6}$$

where $col(H,i) = \sum_{q=1}^{n} A_{qi} col(\dot{H}, q)$ denotes the $i-$th column of H, i.e., the label confidence vector for the $i-$th sample, which is just the linear combination of all columns of $\dot{H}$. Note that L_s is the normalized graph Laplacian of the visual context $L_s = I_{n \times n} - \widetilde{S}$. Since $AL_sA^\top = A(I_{n\times n} - A^\top\Lambda^{-1}A)A^\top = AA^\top - AA^\top\Lambda^{-1}AA^\top$, we can compute $AL_sA^\top \in R^{n \times n}$ efficiently in Eq. (6), and thus avoid computing the large matrix L_s directly.

At the same time, to capture the semantic context, we also define the weight matrix W to measure the inter-label correlations. W is an $m \times m$ symmetric matrix and its entry can be defined as the harmonic mean of the empirical conditional probabilities:

$$W_{st} = \frac{p(t|s)p(s|t)}{(p(t|s) + p(s|t))/2}. \tag{7}$$

where the empirical conditional probabilities $p(t|s) = \frac{\sum_{i=1}^{l} \mathbf{y}_s^i \mathbf{y}_t^i}{\sum_{i=1}^{l} \mathbf{y}_s^i}$ and $p(s|t) = \frac{\sum_{i=1}^{l} \mathbf{y}_s^i \mathbf{y}_t^i}{\sum_{i=1}^{l} \mathbf{y}_t^i}$. The larger the weight is, the stronger the semantic relation is; so, the weight W_{st} measures the correlations between concepts c_s and c_t. Let $\widetilde{W} = D_w^{-\frac{1}{2}} W D_w^{-\frac{1}{2}}$, where D_w is an $m \times m$ diagonal matrix with $D_w(t,t) = \sum_{s=1}^{m} W_{st}$. Since each row of $\dot{H}$ can be viewed as the feature vector of a certain concept, we can achieve the goal that strongly correlated concepts should have similar feature vector by:

$$\begin{aligned} \min_{\dot{H}} \quad & \frac{1}{2} \sum_{s,t} \widetilde{W}_{st} \big\|row(\dot{H}, s) - row(\dot{H}, t)\big\|^2 \\ & = Tr\big(\dot{H}^\top L_w \dot{H}\big) \end{aligned} \tag{8}$$

where $row(\dot{H}, s)$ denotes the $s-$th row of $\dot{H}$, L_w is the normalized graph Laplacian of the semantic context $L_w = I_{m \times m} - \widetilde{W}$.

It should be pointed out that the correlations between concepts derived by the empirical conditional probabilities in Eq. (7) is simple and effective if the labeled samples are sufficient. If the available training images are weakly labeled, there is lack of sufficient training samples and the empirical conditional probabilities might not be estimated correctly, thus the semantic graph with the edge weights Eq. (7) is not expected to capture the semantic context relationship well. To address this problem, we should learn the semantic context and the label confidence vectors, simultaneously. By incorporating various information (including visual context and semantic context) in a single framework, the proposed model takes the formulation as follows:

$$\begin{aligned} \min_{\dot{H}, L_w} f = & Tr\big((Y - \dot{H}A_l)^\top (Y - \dot{H}A_l)\big) + \\ & \theta_1 Tr\big(\dot{H}AL_sA^\top\dot{H}^\top\big) + \theta_2 Tr\big(\dot{H}^\top L_w \dot{H}\big) \end{aligned} \tag{9}$$

where θ_1 and θ_2 are the trade-off parameters.

The cost function (9) can be minimized by updating $\dot{H}$ and L_w alternatively. We derive the gradients of the above cost function with respect to $\dot{H}$ and L_w, respectively:

$$\begin{aligned} \frac{\partial f}{\partial \dot{H}} &= 2(\dot{H}A_lA_l^\top - YA_l^\top + \theta_1 \dot{H}AL_sA^\top + \theta_2 L_w \dot{H}) \\ \frac{\partial f}{\partial L_w} &= \theta_2 \dot{H}\dot{H}^\top \end{aligned} \tag{10}$$

Now the optimal $\dot{H}$ and L_w can be obtained via the iterative alternating updating procedure as follows:

$$\begin{aligned} \dot{H}^t &= \dot{H}^{t-1} - \alpha \frac{\partial f}{\partial \dot{H}}(\dot{H}^{t-1}, L_w^{t-1}) \\ L_w^t &= L_w^{t-1} - \beta \frac{\partial f}{\partial L_w}(\dot{H}^{t-1}, L_w^{t-1}) \end{aligned} \tag{11}$$

where α and β ($0 < \alpha, \beta < 1$) are both the step sizes for gradient search.

To get the initial label confidence matrix $\dot{H}^0$, we can employ SVM or other existing image annotation algorithms like TagProp [9] by using the labeled images as training samples. Based on Eq. (7), we also initialize the normalized graph Laplacian of the semantic context L_w^0.

Once the optimal graph Laplacian L_w^{opt} is learned, the optimal semantic context $\widetilde{W}^{opt}$ is computed straightforward: $\widetilde{W}^{opt} = I_{m\times m} - L_w^{opt}$. Based on the learned optimal $\dot{H}$, the label confidence vector for each image can be obtained according to Eq. (3). By choosing a threshold for each component of the label confidence vector, we can predict the label vectors for each image easily.

4. EXPERIMENTS

We evaluate our method on the real-world image dataset NUS-WIDE [5] which is a challenging collection of web images from Flickr comprising 269,648 images with over 5,000 user-provided tags. Since the ground-truth of 81 concepts for the entire dataset can be used for evaluation, we focus on the 81 concepts in experiments. As in [3], two image pools are constructed from the entire dataset: the pool of labeled images is comprised of 161,789 images while the rest are used for the pool of unlabeled ones. For each image, we first extract two types of visual features: 512-Dim GIST and 1024-Dim SIFT. 3000 exemplar images are selected in experiments.

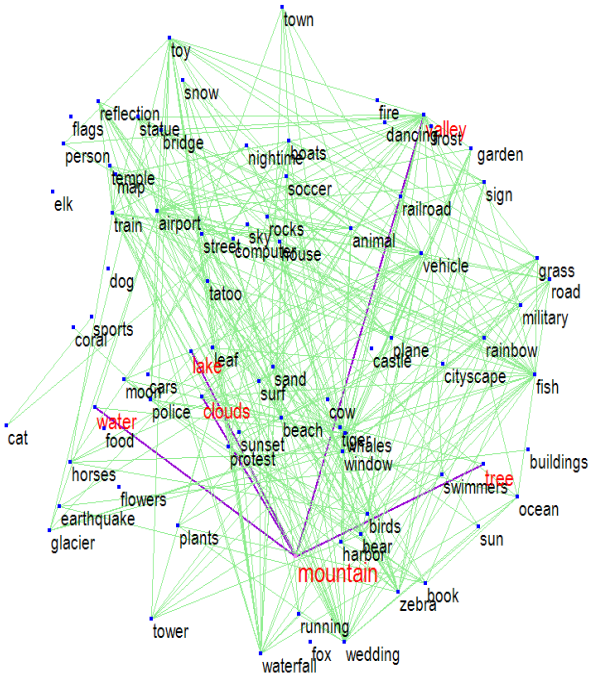

Figure 2: Semantic context learned from NUS-WIDE dataset. Each concept is linked with relevant concepts with larger weights.

Fig.2 shows the semantic context learned from NUS-WIDE image dataset, where each concept is linked with its relevant concepts with larger weights. For some concepts, their inter-concept contexts could be very weak (i.e., having smaller weights), thus it is not necessary for each concept to be linked with all the other concepts. As an example in Fig. 2, the concept *mountain* is more relevant to those concepts *water*, *clouds*, *tree*, *lake*, and *valley* in the NUS-WIDE dataset.

Fig.3 shows the results of our method (Ours) in comparison with the baselines SVM and TagProp [9] in terms of F score for individual concepts on the unlabeled pool of NUS-WIDE images. F score is defined as the harmonic mean of precision and recall:

$$F = \frac{precision * recall}{(precision + recall)/2} \tag{12}$$

As to the baseline SVM, we train one binary SVM for each concept; then 81 SVMs are learned. These SVMs for different concepts are independent because there are no correlations between concepts are leveraged. As observed from the results, our method outperforms the others for most concepts, which demonstrates that our method can effectively learn semantic context from image dataset, and sufficiently leverage neighborhood context, visual context, and semantic context to improve the annotation performance.

5. CONCLUSIONS

In this paper a novel method is developed to predict label vectors for web images by learning semantic context with large-scale weakly-labeled image dataset in a semi-supervised framework. To address the scalability issue, clustering technique is employed to obtain a small number of exemplar images which cover the whole data cloud. The neighborhood contexts are preserved when mapping images from the visual space to the semantic space. The label vector for each image is estimated as a local combination of the exemplar label vectors. By investigating the label confidence matrix for image exemplars from column and row views, our method sufficiently leverages visual context, semantic context, and neighborhood consistency in both visual and semantic spaces, which is important to multi-label image annotation task.

6. ACKNOWLEDGMENTS

We would like to thank the anonymous reviewers for their helpful comments. This work was supported in part by the STCSM's Programs (No. 10511500703 and No. 12XD1400900), the NSF of China (No.60903077), and the 973 Program (No.2010CB327906).

7. REFERENCES

[1] M. Belkin, P. Niyogi, and V. Sindhwani. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *JMLR*, 7:2399–2434, 2006.

[2] L. Chen, D. Xu, I. W. Tsang, and J. Luo. Tag-based web photo retrieval improved by batch mode re-tagging. In *CVPR*, 2010.

[3] X. Chen, Y. Mu, S. Yan, and T.-S. Chua. Efficient large-scale image annotation by probabilistic collaborative multi-label propagation. In *ACM MM*, 2010.

[4] M. J. Choi, J. J. Lim, A. Torralba, and A. S. Willsky. Exploiting hierarchical context on a large database of object categories. In *CVPR*, 2010.

[5] T.-S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, and Y.-T. Zheng. Nus-wide: A real-world web image database from national university of singapore. In *CIVR*, 2009.

[6] J. Fan, Y. Shen, N. Zhou, and Y. Gao. Harvesting large-scale weakly-tagged image databases from the web. In *CVPR*, 2010.

[7] S. L. Feng, R. Manmatha, and V. Lavrenko. Multiple bernoulli relevance models for image and video annotation. In *CVPR*, 2004.

[8] R. Fergus, Y. Weiss, and A. Torralba. Semi-supervised learning in gigantic image collections. In *NIPS*, 2010.

[9] M. Guillaumin, T. Mensink, J. Verbeek, and C. Schmid. Tagprop: Discriminative metric learning in nearest neighbor models for image auto-annotation. In *ICCV*, 2009.

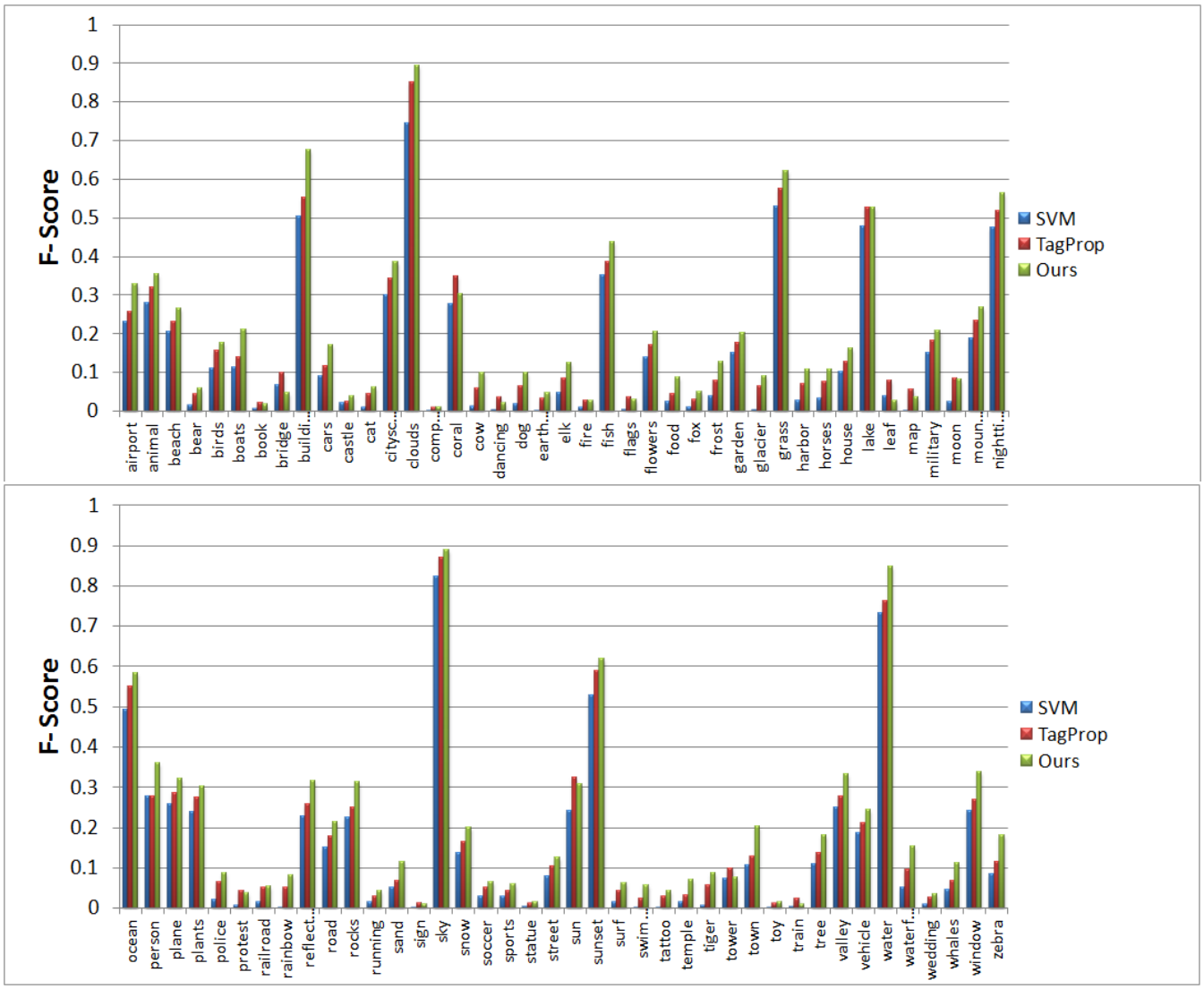

Figure 3: The results of our method in comparison with the baselines SVM and TagProp in terms of F score for individual concepts on NUS-WIDE.

[10] D. Liu, X.-S. Hua, M. Wang, and H.-J. Zhang. Image retagging. In *ACM MM*, 2010.

[11] D. Liu, S. Yan, Y. Rui, and H.-J. Zhang. Unified tag analysis with multi-edge graph. In *ACM MM*, 2010.

[12] W. Liu, J. He, and S.-F. Chang. Large graph construction for scalable semi-supervised learning. In *ICML*, 2010.

[13] A. Makadia, V. Pavlovic, and S. Kumar. A new baseline for image annotation. In *ECCV*, 2008.

[14] S. T. Roweis and L. K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science, vol. 290, 2000*.

[15] X. Rui, M. Li, Z. Li, W.-Y. Ma, and N. Yu. Bipartite graph reinforcement model for web image annotation. In *ACM MM*, 2007.

[16] J. Tang, S. Yan, R. Hong, G.-J. Qi, and T.-S. Chua. Inferring semantic concepts from community contributed images and noisy tags. In *ACM MM*, 2009.

[17] C. Wang, F. Jing, L. Zhang, and H.-J. Zhang. Content-based image annotation refinement. In *CVPR*, 2007.

[18] L. Wu, X.-S. Hua, N. Yu, W.-Y. Ma, and S. Li. Flickr distance. In *ACM MM*, 2008.

[19] Z.-J. Zha, X.-S. Hua, T. Mei, J. Wang, G.-J. Qi, and Z. Wang. Joint multi-label multi-instance learning for image classification. In *CVPR*, 2008.

[20] Z.-J. Zha, T. Mei, J. Wang, Z. Wang, and X.-S. Hua. Graph-based semi-supervised learning with multiple labels. *J. Vis. Commun. Image R., 2009*.

[21] W. Zhang, Y. Lu, X. Xue, and J. Fan. Automatic image annotation with weakly labeled dataset. In *ACM MM*, 2011.

[22] G. Zhu, S. Yan, and Y. Ma. Image tag refinement towards low-rank, content-tag prior and error sparsity. In *ACM MM*, 2010.

[23] X. Zhu, Z. Ghahramani, and J. Lafferty. Semi-supervised learning using gaussian fields and harmonic functions. In *ICML*, 2003.