

Image Annotation with Weak Labels

Feng Tian^{1,2,*} and Xukun Shen¹

¹ State Key Laboratory of Virtual Reality Technology and Systems,
Beihang University, 100191 Beijing, China

² School of Computer and Information Technology,
Northeast Petroleum University, 163318 Daqing, China
{tianfeng,xkshen}@vr1ab.buaa.edu.cn

Abstract. In this paper, we address the problem of image annotation when the given labels of training image are incomplete, inaccurate, and unevenly distributed, in the form of weak labels, which is frequently encountered when dealing with large scale web image training set. We introduce a progressive semantic neighborhood learning approach that explicitly addresses the challenge of learning from weakly labeled image by searching image's semantic consistent neighborhood. Neighbors in image's semantic consistent neighborhood have global similarity, partial correlation, conceptual similarity along with semantic balance. We also present an efficient label inference algorithm to handle noise by minimizing the neighborhood reconstruction error. Experiments with different data sets show that the proposed framework is more effective than the state-of-the-art algorithms in dealing with weakly labeled datasets.





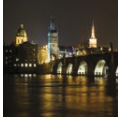
Keywords: label set relevance, web image annotation, image label.

1 Introduction

Traditional image annotation studies, a basic assumption is that all the proper labels of every training image are given and correct. In real environment, this assumption hardly holds since getting all the proper labels is usually expensive, time consuming and people usually add a few labels, rather than an exhaustive list of relevant terms. Moreover, not all of the labels are relevant to the image content ,for example, images labeled with "car" might be taken from a car, rather than depicting one. It is evident that this scenario is quite different from the classic image annotation setting where all proper labels for training data are assumed to be given. Images in benchmark set are also usually weakly labeled(showed in Table 1). Meanwhile, large variations in the frequency of different labels can reduce the performance of the labeling method on the low-frequency labels. E.g., in an experiment on the Corel5K dataset, we found that for the 20% least frequent labels, JEC [1] achieves an F-score of 19.7%, whereas it gives reasonably good performance for the 20% most frequent labels with F-score being 50.6%. In this work, the meaning of the terminology "weak

* Corresponding author.This work is supported by Scientific Research Fund of Heilongjiang Provincial Education Department(NO:12511011,12521055).

Table 1. Weak label image (the missing labels are highlighted by bold font, the content unrelated labels are italicized)

				
field horses mare	fence mountain	<i>travel vacation</i>	bear river	<i>czech</i> bridge
foals	range sky	Nile sky	reflection water	<i>charles lights</i>
tree	airplane	sailboat sea	black	night

labels” is threefold: (1) the given labels may be incomplete, namely only a subset of labels are attached to images according to the ground truth; (2) even for the labels provided, there may be noisy labels; (3) there is large variations in the frequency of different labels (semantic imbalance). Image annotation from weakly labeled dataset is important since weakly-labeled problems are prevalent in the popular datasets as well as real-world environment. In [2], the authors showed performance improvement where for each training instance, only one of its class assignments is correct. In [3], a hybrid model framework for utilizing partially labeled data that integrates a generative topic model for image appearance with discriminative label prediction is explored. In [4], the author focuses on removing false class assignments for training set. In [5], the author proposed ranking based multi-label learning to learn from incompletely data. Our work is more comprehensive and address a more realistic and challenging scenario where the datasets seriously suffer from weakly labeled issues.

2 Our Approach

Based on the idea that negative impact of the weak label can be reduced under the guidance of neighbors, the training image’s labels are replenished by minimizing the label’s weighted error function, then ”semantic balanced neighborhood” is set up based on the replenishing labels to address the large differences in these label’s frequency. Linear metric embedded in multiple label information is learned to obtain the consistency of distance measure and image semantic. Then the images’ partial correlation is obtained by image’s nonnegative sparse linear combination between neighbors. The neighbors in the final neighborhood have higher global similarity, partial correlation and conceptual similarity along with semantic balance. Label prediction is performed in the neighborhood by minimizing label’s reconstruction error loss, and noise labels are handled by two regulation terms.

2.1 Semantic Balanced Neighborhood

By semantic balanced neighborhood (short for BN), we mean, for a given image, there should not have large differences in the frequency of different labels in

it's neighborhood. Considering labels for training image may be incomplete, we replenish missing labels firstly. Denote vocabulary as $C = \{c_1, c_2, \dots, c_q\}$ and training set $L = \{(x_1, y_1), \dots, (x_l, y_l)\}$, m represents the dimensionality of features, $y_i = (y_{i1}, \dots, y_{iq}) \in \{0, 1\}^q$ is the corresponding label vector, $y_{ij} = 1$ if the i -th image has the j -th label and $y_{ij} = 0$ otherwise. $Y = [y_1, \dots, y_l]^T$ be the corresponding label indicator matrix. We want to learn a replenished function $f : L \rightarrow R^q$ where $f_i = [f_{i1}, f_{i2}, \dots, f_{iq}]^T$, f_{ij} denotes the value of function output of i -th image, and we use matrix $F = [f_1, f_2, \dots, f_l]$ to present the replenished label matrix. The error function is $E(f) = E_1(f) + \lambda E_2(f)$, where $\lambda \geq 0$ is a controlling parameter. Thus, the optimization problem is:

$$\min_f \left\{ \frac{1}{2} \sum_{j=1}^q \sum_{i=1}^l u_{ij} (y_{ij} - f_{ij})^2 + \frac{1}{2} \lambda \sum_{i=1}^l \sum_{j=1}^l w_{ij} \left\| \frac{f_i}{\sqrt{d_i}} - \frac{f_j}{\sqrt{d_j}} \right\|^2 \right\}$$

$E_1(f)$ represents the weighted error function, u_{ij} represents the weight between sample x_i and j -th label, $u_{ij} = 1$ if $y_{ij} = 1$, and τ otherwise ($0 \leq \tau \leq 1$). Minimizing $E_1(f)$ is equivalent to requiring the output of f is similar to the original labels and can replenish the missing labels. Minimizing the second term is equivalent to requiring the smoothness output of f on each sample's neighbor according to their similarity. The approximate optimal solution can be derived by least squares. Based on the replenished labels, image's BN is constructed as follows. Let $L_i \subseteq L$ ($\forall i \in \{1, 2, \dots, q\}$) be the subset of training data that contains all the images annotated with the label c_i , we consider it as a semantic group. Given an image x , from each semantic group we pick k_2 images that are most similar to x and form corresponding sets $L_{x,i} \subseteq L_i$. Thus, each $L_{x,i}$ contains images that are most informative in predicting the probability of the label c_i for x . We merge them all to form the semantic balanced neighborhood as $BN(x) = \{L_{x,1} \cup \dots \cup L_{x,q}\}$. It can be easily noted that in $BN(x)$, each label appears (at least) k_2 times, thus addressing the semantic imbalance issue.

2.2 Semantic Consistent Neighborhood

By semantic consistent neighborhood (short for CN), we mean, the neighbors should have both the global similarity and partial correlation along with conceptual similarity. We select the partial correlated neighbors in target image's BN by sparse representation. Note that, from signal reconstruction point of view, when target signal is reconstructed from signals in different subspace (semantic subspace), the reconstruction coefficients have lost their physical meaning. It is obvious that we cannot guarantee that sample's in BN are all semantically similar, since image pair's semantic similarity depends on the corresponding label set instead of single label. As shown in figure1(a), x_p 's semantically similar neighbors (denoted by circle) and neighbors that are semantically dissimilar neighbors (denoted by square) are all in x_p 's BN. If semantically dissimilar neighbors lie outside smaller radius with a margin of at least one unit distance, as shown in figure2(b), then we can reconstruct x_p by neighborhood (b). Let a and

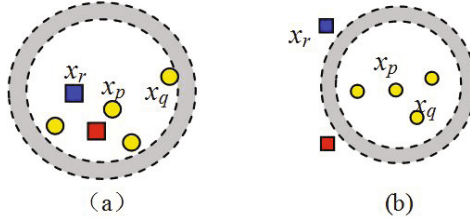


Fig. 1. Schematic illustration of one input's neighborhood

b be two training images, and each represented by n features $\{f_A^1, \dots, f_A^n\}$ and $\{f_B^1, \dots, f_B^n\}$, where $\sum m_i = m$.

$$\tilde{d}(a, b) = \sum_{i=1}^n w(i) \sum_{j=1}^{m_i} u^i(j) \cdot dist_{AB}^i(j) \tag{1}$$

u^i and w are usually taken as a non-negative normalized unit vector, u^i can be assigned appropriate weights to individual dimensions of a feature vector in the feature space, w is to optimally combine multiple feature distances. Given an image x_p , along with its label vector y_p , we want to learn weights such that its target neighbor x_q from the semantic groups $\{L_{x_{p,r}}\}_r$ are pulled closer, x_r from the remaining semantic groups are pushed far. That is, minimize the error function:

$$\begin{aligned} & argmin_{w,u} \sum_{pq} \eta_{pq} \lambda_{pq} \tilde{d}(x_p, x_q) + \\ & \mu \sum_{pqr} \eta_{pq} (1 - \lambda_{pr}) [1 + \tilde{d}(x_p, x_q) - \tilde{d}(x_p, x_r)]_+ \end{aligned}$$

Here, μ is the controlling parameters, $[z]_+ = max(0, z)$ is the hinge loss, λ_{pq} and λ_{pr} scale the error loss depending on the overlap between the label sets of images. We solve it by alternatively using stochastic sub-gradient descent and projection steps (similar to Pegasos [6]) to obtain an approximate optimal solution of w and u^i . Then, given image x_i , we find it's k nearest neighbor by Equation (1) to construct x_i 's local overcomplete dictionary, where $i_p \in \{1, \dots, l\}$, $p \in \{1, \dots, k\}$. α_i is the reconstruction coefficients vector for x_i . Note that negative coefficient has not explicit meaning to describe semantic, so we reformulate the reconstruction

relationship as $x_i = B_i \alpha_i + \zeta$, where $\alpha_i(p) \geq 0$ and $\sum_{p=1}^k \alpha_i(p) = 1$. Let non-

negative term ζ^+ , noise term $\zeta = \zeta^+ - \zeta^-$, $|\zeta| = \zeta^+ + \zeta^-$. Then we can solve $\min_{\alpha_i} \lambda \|\alpha_i\|_1 + \frac{1}{2} \|x_i - B_i \alpha_i\|_2^2$ s.t. $\alpha_i \geq 0$ where $x_i = [x_i \ 1]^T$, $\alpha_i = [\alpha_i \ \zeta^+ \ \zeta^-]$

$B_i = \begin{bmatrix} B_i & I_m & -I_m \\ E_{1 \times k} & 0_{1 \times m} & 0_{1 \times m} \end{bmatrix}$, the controlling term $\lambda = 2\|B_i x_i\|_\infty$, the problem can be solved efficiently using L1 optimization toolbox like YALL. Then, x_i is represented by a sparse linear combination of it's neighbors, and it's semantic consistent neighborhood(CN) is composed by the neighbors x_{i_p} where $a_i(p) > 0$. Let $C = [c_{ij}]$ denotes neighborhood weight matrix, where $c_{ij} = \alpha_i(p)$.

2.3 Label Inference in CN

Let $f = [f_L f_U]^T$ be label matrix of all samples, where f_L represents the training set's label matrix, f_U represents the unlabeled ones' label matrix(initialized by zero). Assuming that each image's label vector can be reconstructed by it's neighbors in it's CN, while the reconstruction coefficients are the same as their visual reconstruction coefficient. Thus we can predict the labels of the unlabeled samples by the weight in neighborhood matrix C . This prediction is based on the assumption that the weight c_{ij} reflects the likelihood for sample x_i to have the same label as sample x_j . So the labels of the unlabeled samples can be inferred by minimizing label reconstruction error as follows:

$$E(f) = \sum_{i=1}^n \|f_i - \sum_{j \neq i} c_{ij} f_j\|^2, s.t. f_i = y_i \quad (2)$$

where y_i is the replenished label vector of x_i . We use generalized minimum residual method (GMRES [7]) to obtain an approximated solution. As aforementioned, the associated labels are often incomplete and imprecise, so the training labels cannot be fixed during the inference process as in Equation (2) should be refined. However, the training labels should be consistent with the original labels to a certain extent. So the optimization target should be

$$\min_f \left\{ \|f - Cf\|^2 + \lambda_1 \|f_L - \hat{f}_L\|^2 + \lambda_2 \|\hat{f}_L - Y\|_1 \right\}$$

where f_L is the training images labels that are propagated, and \hat{f}_L denotes the ideal label vector of the training images. The first term of this formula is the same as in Equation (2). The second term enforces the ideal labels of the training images to be consistent with the labels propagated. The third term constrains that only a limited number of labels are noisy or imprecise.

3 Experiments

We validate the effectiveness of our proposed approach on IAPR-TC12,ESPGAME and FLICKR2.5M datasets. Each image is annotated with

Table 2. Experimental results on four dataset

Method	IAPR				ESP				FLI		
	P	R	F1	N+ P	R	F1	N+ P	R	F1	N+	
SML[8]	0.21	0.23	0.220	201 0.16	0.17	0.165	195 0.15	0.16	0.155	278	
JEC [1]	0.28	0.29	0.285	250 0.22	0.25	0.234	224 0.20	0.21	0.205	355	
Tagprop(ml) [9]	0.48	0.25	0.329	227 0.49	0.20	0.284	213 0.43	0.17	0.244	363	
Tagprop(ml+s)[9]	0.46	0.35	0.398	266 0.39	0.27	0.319	239 0.34	0.25	0.288	403	
GS [10]	0.32	0.29	0.304	252 0.36	0.24	0.288	226 0.29	0.22	0.250	375	
SNLWL	0.52	0.37	0.432	276 0.51	0.30	0.378	249 0.48	0.29	0.362	427	

5 most relevant keywords. Results are summarized in Table 2. From the results we find that our method significantly improve over the current state-of-the-art. On ESP-GAME image dataset, we achieve 30.8% and 11.1% performance improvement in terms of precision and recall. On FLICKR, we achieve 41.2% and 24% performance improvement in terms of precision and recall.

4 Conclusion

Image annotation with weakly labeled images is important since weakly-labeled problems are prevalent in the popular web datasets as well as real-world environment. Experiments show that the proposed framework is more effective than state-of-the-art over web image dataset.

References

1. Makadia, A., Pavlovic, V., Kumar, S.: A new baseline for image annotation. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008, Part III. LNCS, vol. 5304, pp. 316–329. Springer, Heidelberg (2008)
2. Nguyen, N.: Classification with partial labels. In: ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 551–559. ACM Press, New York (2008)
3. He, X.: Learning hybrid models for image annotation with partially labeled data. In: Conference on Neural Information Processing Systems, pp. 625–632 (2008)
4. Fan, J.: Harvesting large-scale weakly-tagged image databases from the web. In: International Conference on Computer Vision, pp. 802–809. IEEE Computer Society Press, Los Alamitos (2010)
5. Bucak, S.: Multi-label learning with incomplete class assignments. In: International Conference on Computer Vision, pp. 2801–2808. IEEE Computer Society Press, Los Alamitos (2011)
6. Shwartz, S.: Pegasos: Primal estimated sub-gradient solver for svm. In: International Conference on Machine Learning, pp. 807–814 (2007)
7. Saad, Y.: Gmres: A generalized minimal residual algorithm for solving nonsymmetric linear systems. In: International Conference on Computer Vision; SIAM Journal on Scientific and Statistical Computing 7, 856–869 (1986)
8. Carneiro, G.: Supervised learning of semantic classes for image tagging and retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 29, 394–410 (2007)
9. Guillaumin, M.: Tagprop: Discriminative Metric Learning in Nearest Neighbor Models for Image Auto-tagging. In: International Conference on Computer Vision, pp. 309–316. IEEE Computer Society Press, Los Alamitos (2009)
10. Zhang, S.: Automatic image annotation using group sparsity. In: International Conference on Computer Vision, pp. 3312–3319. IEEE Computer Society Press, Los Alamitos (2010)