IMAGE ANNOTATION VIA LEARNING THE IMAGE-LABEL INTERRELATIONS

Yonghao He, Jian Wang, Shiming Xiang, Chunhong Pan

NLPR, Institute of Automation, Chinese Academy of Sciences, Beijing, China
yhhe@nlpr.ia.ac.cn, jwang5209@ia.ac.cn, {smxiang, chpan}@nlpr.ia.ac.cn

ABSTRACT

The goal of image annotation is to automatically assign meaningful and content-related labels to the digital images by using machines. It is beneficial to image search and image sharing in social networks. Various methods for image annotation are proposed in last decade and they have gained much progress. However, most of them are not precise and fast enough for real-world applications. In this paper, we propose a novel and fast image annotation method via learning the image-label interrelation. The main idea of the proposed method is to predict labels by linearly propagating the label information through the image-label interrelation and the image similarities. Thus, we propose a model based on the regression between the label groundtruth and the propagated label information to learn the image-label interrelation. In addition, a label-biased regularization is integrated into our model to learn more effective and meaningful image-label interrelation. Finally, our model can be solved in closed form, therefore it achieves a fast learning process. Experimental results on three benchmark datasets demonstrate that our method shows the comparable performance with state-of-the-art methods and has faster learning time.

Index Terms— Image Annotation, Image-label Interrelation, Closed Form Solution, Fast Annotation

1. INTRODUCTION

Annotating images by using machines is a very significant mission. The tremendous growth of web images puts forward some urgent issues. First, in traditional image retrieval systems, matching keywords is the most widely used technique in which the precondition is that each image should be annotated. Thus, annotating such great amount of unlabeled images is inevitable. Second, one reason of the prosperity of social networks is mainly driven by the large number of pictures uploaded by users. In most cases, users are requested to attach meaningful labels to images so as to be offered more services such as efficient picture sharing and interests group recommendation. So users have to confront such tedious annotating process. To address above issues, automatic image annotation is one of the optimal choices for avoiding time-consuming and monotonous work for humans.

Extensive studies have been made for image annotation. According to the different points of interest, these methods can be roughly divided into two categories: classification-based and modeling-based. The classification-based methods [1, 2, 3] view concepts (labels) as classes and train classifiers for each concept. The drawback of this kind of methods lies in that image annotation is about multi-label predictions rather than isolated one-vs-all classifications and the interactions between the images and the labels are ignored. Most methods [4, 5, 6, 7, 8, 9, 10, 11] fall into the second category. Feng *et al.* [6] and Jeon *et al.* [7] try to annotate images from the perspective of probability, and they model the joint distributions of the image features and the labels. The authors of [4] present a multi-label sparse coding framework in which the main idea is to propagate label information from training images to query image by sparse coding coefficients.

Recently, algorithms in [10, 11] show state-of-the-art performances on image annotation. Guillaumin *et al.* [10] propose an algorithm based on weighted nearest neighbor model. It allows the integration of metric learning in the procedure of maximizing the log-likelihood, and a word specific sigmoidal modulation is also incorporated to promote the recall of rare labels. Additionally, multiple types of features can be easily used in this algorithm to compute the image similarities. However this method takes too much time on training, which prevents it from being applied to real-world applications. To overcome the drawback of time-consuming in [10], Chen *et al.* [11] design a fast annotation algorithm which maps the image features to the label space. A novel marginalized blank-out regularization to enrich tags is proposed to make the projection matrix more effective. FastTag [11] greatly decreases computational time by solving the convex optimization in closed form. However, when utilizing multiple features, FastTag has to do dimensionality reduction to keep the learning procedure fast, which may cost much time.

In this paper, we propose a novel and fast annotation method via learning the image-label interrelation, which contains two advantages—multiple features can be easily used without dimensionality reduction and the final objective func-

This work was supported in part by the National Basic Research Program of China under Grant 2012CB316304, the National Natural Science Foundation of China under Grants 61272331 and 61272049, and the Beijing Natural Science Foundation (No. 4132075).

tion can be elegantly solved in closed form, resulting in a fast learning process. Inspired by the work in [4], we recognize that propagating label information through a simple image-label interrelation matrix (only contains 0s and 1s entries, 0 for label absence and 1 for label presence) and the image similarities can provide promising predictions. Hence, a model based on regression, which minimizes the distance between the label groundtruth and the propagated predictions, is proposed to learn an effective image-label interrelation matrix. Moreover, we add a label-biased regularization that enlarges the entries of label presence and suppresses the entries of label absence to guide the model to learn a more meaningful image-label interrelation matrix. The image similarities are precomputed using multiple features and this process needs no dimensionality reduction. The resulting objective function can be solved in closed form. Eventually, our method achieves comparable performance with state-of-the-art methods [10, 11] using less learning time.

2. THE PROPOSED METHOD

2.1. Motivation

In this section, we first introduce the algorithm in [4]. In [4], complicated feature representation and dimensionality reduction are first developed; then each image to be annotated is sparsely reconstructed by all training images; finally the reconstruction coefficients and the label matrix of training images are used for predicting labels:

$$\mathbf{c} = \mathbf{C}\boldsymbol{\alpha}, \tag{1}$$

where $\boldsymbol{\alpha}$ is the reconstruction coefficient vector, $\mathbf{C}$ is the label matrix of the training images with 0 or 1 entries, and the top labels with the largest values in $\mathbf{c}$ are considered as the final annotations. Essentially, the reconstruction coefficients $\boldsymbol{\alpha}$ can be viewed as the similarities between the test image and all training images, and the label matrix $\mathbf{C}$ can be deemed as a simple image-label interrelation matrix that only captures the existence and presence relationships. Thus, the label can be predicted by linearly propagating the label information through the image-label interrelation and the image similarities. From this point, we are inspired to learn better image-label interrelation that can represent richer relationships between the images and the labels.

2.2. Learning the Image-Label Interrelation

Notation. Let T denote the size of vocabulary. The training set is $\mathcal{D} = \{(\mathbf{x}_1, \mathbf{y}_1), (\mathbf{x}_2, \mathbf{y}_2), \cdots, (\mathbf{x}_N, \mathbf{y}_N)\}$, where $\mathbf{x}_i = \{\mathbf{x}_i^1, \mathbf{x}_i^2, \cdots, \mathbf{x}_i^K\}$ is the set of K types of features for the i-th image, in which $\mathbf{x}_i^k$ is a certain feature vector, and $\mathbf{y}_i \in \mathcal{R}^T$ is the corresponding label indicator vector with 0 or 1 entries. Our goal is to learn the image-label interrelation matrix $\mathbf{M} \in \mathcal{R}^{T \times N}$. Each row in $\mathbf{M}$ reflects the relationships between a label and all images, and each column in $\mathbf{M}$ represents the correlations between an image and all labels.

The discussion in Section 2.1 indicates that the image similarities and the simple label matrix are potentially effective for attaching correct labels. Thus, a straightforward way to learn the image-label interrelation matrix is to solve the following linear regression:

$$\min_{\mathbf{M}} \frac{1}{N} \sum_{i=1}^{N} \|\mathbf{y}_i - \mathbf{M}\mathbf{s}_i\|_2^2, \tag{2}$$

where $\mathbf{s}_i \in \mathcal{R}^N$ collects the similarities between the i-th image and all training images. Moreover, the similarity between any two images is computed by Gaussian weighting function:

$$s(i,j) = \exp\left(-\frac{\frac{1}{K}\sum_{k=1}^{K} d_k(\mathbf{x}_i^k, \mathbf{x}_j^k)}{\sigma}\right), \tag{3}$$

where $d_k(\cdot)$ is a feature-related measurement including Euclidean distance, l_1 distance or χ^2 distance, and σ is the radius of Gaussian function. Compared to [11], using similarities instead of the raw visual features brings some advantages: 1) employing more features easily; 2) avoiding dimensionality reduction; 3) weakening the influence of noisy features.

The objective function in (2) is formulated in form of least squares without any regularization terms. To obtain a better image-label interrelation matrix, we consider to utilize meaningful regularizations to guide the model to learn a more effective $\mathbf{M}$.

Although $\mathbf{M}$ can capture more information than that in label matrix $\mathbf{C}$, $\mathbf{C}$ still reminds us that entries with label presence should be naturally larger than those with label absence in $\mathbf{M}$. By doing this, the precision and the recall can be promoted. Accordingly, we mathematically present this intuition as follows:

$$\max_{\mathbf{M}} \left(\sum_{i=1}^{N} \mathbf{y}_i^T \mathbf{M}_{:i} - \sum_{i=1}^{N} \hat{\mathbf{y}}_i^T \mathbf{M}_{:i}\right), \tag{4}$$

where $\mathbf{M}_{:i}$ is the i-th column of $\mathbf{M}$, and $\hat{\mathbf{y}}_i =\sim \mathbf{y}_i$, which means that 0s are changed to 1s and 1s are changed to 0s in $\mathbf{y}_i$. This regularization strongly suggests that larger values should be assigned to the entries with label presence and smaller values should be assigned to entries with label absence.

Integrating the label-biased regularization into the objective function in (2), we obtain:

$$\begin{aligned} \min_{\mathbf{M}} f(\mathbf{M}) \quad = \quad & \frac{1}{N} \sum_{i=1}^{N} \|\mathbf{y}_i - \mathbf{M}\mathbf{s}_i\|_2^2 + \beta\Bigg(\sum_{i=1}^{N} \hat{\mathbf{y}}_i^T \mathbf{M}_{:i} \\ & - \sum_{i=1}^{N} \mathbf{y}_i^T \mathbf{M}_{:i}\Bigg) + \lambda \|\mathbf{M}\|_F^2, \end{aligned} \tag{5}$$

where β and λ are trade-off parameters, and the last term $\|\mathbf{M}\|_F$ is Frobenius norm of $\mathbf{M}$ that is to guarantee an available solution and avoid over-fitting. Then we rewrite (5) to a more compact matrix form:

Table 1. Annotation performances on three datasets. Methods include state-of-the-art ones [11, 10] and some other reported algorithms [1, 4, 8, 6, 9]. The performance of our method is in the last line. ("-" means not reported.)

	Corel5K				IAPRTC-12				ESP game			
Method	**P**	**R**	**F1**	$\mathbf{N^+}$	**P**	**R**	**F1**	$\mathbf{N^+}$	**P**	**R**	**F1**	$\mathbf{N^+}$
SML [1]	0.23	0.29	0.26	137	-	-	-	-	-	-	-	-
MSC [4]	0.25	0.32	0.28	136	-	-	-	-	-	-	-	-
TGLM [8]	0.25	0.29	0.27	131	-	-	-	-	-	-	-	-
MBRM [6]	0.24	0.25	0.24	122	0.24	0.23	0.23	223	0.18	0.19	0.18	209
JEC [9]	0.27	0.32	0.29	139	0.29	0.19	0.23	211	0.24	0.19	0.21	222
TagProp [10]	0.33	0.42	0.37	160	0.45	0.34	0.39	260	0.39	0.27	0.32	238
FastTag [11]	0.32	0.43	0.37	166	0.47	0.26	0.34	280	0.46	0.22	0.30	247
Our method	0.31	0.42	0.36	158	0.43	0.33	0.37	257	0.30	0.26	0.28	238

$$\min_{\mathbf{M}} f(\mathbf{M}) = \frac{1}{N}\|\mathbf{Y}-\mathbf{MS}\|_F^2 + \beta\Big(tr(\hat{\mathbf{Y}}^T\mathbf{M}) - tr(\mathbf{Y}^T\mathbf{M})\Big) + \lambda\|\mathbf{M}\|_F^2, \quad (6)$$

where $\mathbf{Y} = [\mathbf{y}_1\ \mathbf{y}_2\ \cdots\ \mathbf{y}_N] \in \mathcal{R}^{T\times N}$, $\hat{\mathbf{Y}} = [\hat{\mathbf{y}}_1\ \hat{\mathbf{y}}_2\ \cdots\ \hat{\mathbf{y}}_N] \in \mathcal{R}^{T\times N}$, $\mathbf{S} = [\mathbf{s}_1\ \mathbf{s}_2\ \cdots\ \mathbf{s}_N] \in \mathcal{R}^{N\times N}$ and $tr(\cdot)$ denotes the trace of a matrix. Then we take the derivative of $f(\mathbf{M})$ with respect to $\mathbf{M}$ in (6):

$$\frac{\partial f}{\partial \mathbf{M}} = \frac{2}{N}\left(\mathbf{MSS}^T - \mathbf{YS}^T\right) + \beta\left(\hat{\mathbf{Y}} - \mathbf{Y}\right) + 2\lambda\mathbf{M}, \quad (7)$$

and set it to zero, we can obtain:

$$\mathbf{M} = \left(\frac{\beta N}{2}\left(\mathbf{Y} - \hat{\mathbf{Y}}\right) + \mathbf{YS}^T\right)\left(\mathbf{SS}^T + \lambda N\mathbf{I}\right)^{-1}. \quad (8)$$

From Eq. (8), we observe that $\mathbf{M}$ can be solved in closed form. Note that, the main calculation of $\mathbf{M}$ is concerned with the inverse of a matrix ($O(N^3)$). In FastTag, block-coordinate descent based on two closed form solutions and some other optimization stages make the training time longer. The learning procedure of TagProp is based on gradient descent and each iteration involves the whole training set, thus it may cost much more time. It is difficult to estimate the time complexities of FastTag and TagProp. We present the learning time of three methods on three datasets in Section 3.3.

The final prediction process is similar to Eq. (1). First, the similarities $\mathbf{s}_q$ between the query image $\mathbf{x}_q$ and all training images are calculated by Eq. (3). Second, label scores are obtained by computing the following formula:

$$\mathbf{r}_q = \mathbf{Ms}_q. \quad (9)$$

Finally, labels with the largest values in $\mathbf{r}_q$ are selected as the final annotations.

3. EXPERIMENTS

3.1. Datasets and Features

Our experiments are carried out on three benchmark data-sets, namely Corel5K, IAPRTC-12 and ESP game (All the datasets are obtained from the website: http://lear.inrialpes.fr/people/guillaumin/data.php).

Corel5K [5]. The dataset is the most widely used one for image annotation task. There are 5000 images collected from 50 categories with each category including 100 images. Each image is annotated with 1 to 5 labels, and there are 260 words in both training and testing sets.

IAPRTC-12 [12]. The dataset consists of 19627 images from diverse categories, such as sports, people and landscapes. Each image has 5.7 labels on average and the vocabulary size is 291.

ESP game [13]. The dataset consists of 20770 images depicting logos, drawings and personal photos. All images are annotated with 268 distinct words, and 4.7 labels are assigned to each image on average.

For the training/test splits of three datasets, we follow the criterion in [10].

As for the features, we use the same ones that are released by the authors of [10]. Definitely, there are 15 different types of features including global and local features. For global features, there are Gist features [14] and color histograms (RGB, LAB, HSV) with 16 bins in each channel. Well-known SIFT descriptor [15] and robust hue descriptor [16] composing the local features are extracted densely from multi-scale grids or based on the points detected by Harris corner detector [17]. For the feature-related measurement $d_k(\cdot)$ in Eq. (3), we also follow the criterion that is described in [10].

3.2. Results

Widely used measures for image annotation are used: mean precision (P) and mean recall (R) for each label, $F1$ score

Fig. 1. Comparison between the predicted labels by our method and the groundtruth labels on IAPRTC-12. Only top 5 predicted labels are taken. Labels in red mean that they do not appear in groundtruth.

($F1 = \frac{2*P*R}{P+R}$) and N^+ (the number of labels with non-zero recall value). Experimental results are presented in Table 1.

There are three parameters in our algorithm, namely σ, β and λ. Empirically, we found that $\sigma = 0.34$ and $\lambda = 10^{-4}$ can achieve optimal results with fixed β on three datasets, thus σ and λ is stable to the datasets. As for β, here we describe its influences instead of plotting in figures due to the space limitation. For three datasets, β is tuned from 10^{-4} to 0.5. On Corel5K, IAPRTC-12 and ESP game, when $\beta > 0.2$, $\beta > 0.05$ and $\beta > 0.0005$, the optimal performances are obtained and the performances keep stable.

We can make some observations according to Table 1. First, the early proposed algorithms [1, 4, 8, 6, 9] achieve unsatisfactory performances and can hardly be applied to applications. Second, our method significantly outperforms MSC [4] that inspires our method on Corel5K. Specifically, 24% improvement on mean precision, 31% on mean recall and extra 22 N^+ is gained. This firmly demonstrates that the carefully learned image-label interrelations are more effective than the simple presence and absence relationships. Third, TagProp [10] and FastTag [11] show state-of-the-art performances and our method has comparable performance with them on three datasets. Fourth, due to the novel marginalized blank-out regularization, the recently proposed FastTag [11] presents higher scores on mean precision and N^+, but relatively lower on mean recall on IAPRTC-12 and ESP game. Finally, the only unsatisfactory result of our method is the low score on mean precision on ESP game, which is 9% and 16% lower than TagProp [10] and FastTag [11], respectively.

Additionally, we also present the comparison between the annotations, predicted by our method, of some exemplar images and their groundtruth labels on IAPRTC-12 in Fig. 1.

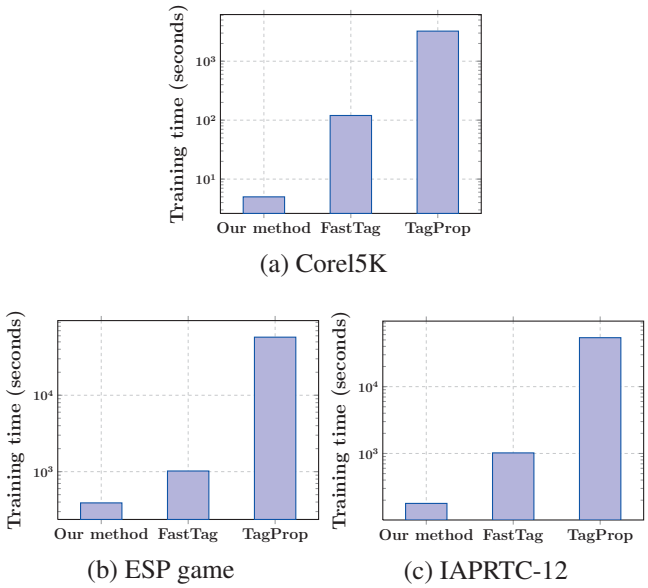

Fig. 2. Training time of three methods on three datasets. The vertical axis represents training time in seconds that is in log scale.

3.3. Computational Efficiency

In this section, we compare the learning time between our method and state-of-the-art methods [11, 10]. For the fair comparison, we directly acquire the reported results in [11] which can avoid some implementation tricks. Although these results are obtained on a desktop with dual 6-core Intel i7 cpus with 2.66Ghz that is different from ours, we still can firmly prove the superiority of the computational efficiency of our method by achieving shorter learning time with a lower hardware configuration—2-core Intel i3 cpus with 3.10Ghz.

Figure 2 illustrates the learning time of our method and other two state-of-the-art methods. Brief analysis of the complexities of three methods is in Section 2.2. In Fig. 2, we can observe that our method is over 100x faster than TagProp and 10x faster than FastTag. So our method is definitely superior to other two method in terms of computational efficiency.

4. CONCLUSIONS

The main idea of the proposed method lies in that the label information can be linearly propagated through the carefully learned image-label interrelation and the image similarities. To this end, we directly model this idea based on the regression between the label groundtruth and the propagated label information. By adding a label-biased regularization, the model gains the ability to learn a more effective and meaningful image-label interrelation matrix. Finally, we obtain a compact objective function and solve it in closed form. The experimental results validate that our method has comparable performance with state-of-the-art methods, and it shows a faster learning process.

5. REFERENCES

[1] G. Carneiro, A. Chan, P. Moreno, and N. Vasconcelos, "Supervised learning of semantic classes for image annotation and retrieval," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 3, pp. 394–410, 2007.

[2] C. Cusano, G. Ciocca, and R. Schettini, "Image annotation using svm," in *Proceedings of Internet imaging V*, 2004, pp. 330–338.

[3] J. Li and J. Wang, "Automatic linguistic indexing of pictures by a statistical modeling approach," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 9, pp. 1075–1088, 2003.

[4] C. Wang, S. Yan, L. Zhang, and H. Zhang, "Multi-label sparse coding for automatic image annotation," in *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, 2009, pp. 1643–1650.

[5] P. Duygulu, K. Barnard, J. de Freitas, and D. Forsyth, "Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary," in *Proceedings of European Conference on Computer Vision*, pp. 97–112. Springer, 2006.

[6] S. Feng, R. Manmatha, and V. Lavrenko, "Multiple bernoulli relevance models for image and video annotation," in *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, 2004, pp. 1002–1009.

[7] J. Jeon, V. Lavrenko, and R. Manmatha, "Automatic image annotation and retrieval using cross-media relevance models," in *Proceedings of ACM SIGIR*, 2003, pp. 119–126.

[8] J. Liu, M. Li, Q. Liu, H. Lu, and S. Ma, "Image annotation via graph learning," *Pattern recognition*, vol. 42, no. 2, pp. 218–228, 2009.

[9] A. Makadia, V. Pavlovic, and S. Kumar, "A new baseline for image annotation," in *Proceedings of European Conference on Computer Vision*, pp. 316–329. 2008.

[10] M. Guillaumin, T. Mensink, J. Verbeek, and C. Schmid, "Tagprop: Discriminative metric learning in nearest neighbor models for image auto-annotation," in *Proceedings of IEEE International Conference on Computer Vision*, 2009, pp. 309–316.

[11] M. Chen, A. Zheng, and K. Weinberger, "Fast image tagging," in *Proceedings of International Conference on Machine Learning*, 2013, pp. 1274–1282.

[12] M. Grubinger, P. Clough, H. Muller, and T. Deselaers, "The iapr tc-12 benchmark: A new evaluation resource for visual information systems," in *International Workshop OntoImage*, 2006, pp. 13–23.

[13] L. Von Ahn and L. Dabbish, "Labeling images with a computer game," in *Proceedings of ACM SIGCHI Conference on Human Factors in Computing Systems*, 2004, pp. 319–326.

[14] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *International Journal of Computer Vision*, vol. 42, no. 3, pp. 145–175, 2001.

[15] D. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[16] J. Van De Weijer and C. Schmid, "Coloring local feature extraction," in *Proceedings of European Conference on Computer Vision*, pp. 334–348. 2006.

[17] C. Harris and M. Stephens, "A combined corner and edge detector.," in *Alvey Vision Conference*, 1988, pp. 147–152.