MULTI-SOURCE IMAGE AUTO-ANNOTATION

Zijia Lin[⋆] *Guiguang Ding*[†] *Mingqing Hu*[‡]

[⋆] Department of Computer Science and Technology, Tsinghua University, Beijing, 100084, China
[†]School of Software, Tsinghua University, Beijing, 100084, China
[‡]Institute of Computing Technology, Chinese Academy of Sciences, Beijing, 100190, China
linzijia07@tsinghua.org.cn dinggg@tsinghua.edu.cn humingqing@ict.ac.cn

ABSTRACT

Though the field of image auto-annotation has been extensively researched, most previous work concentrated on the single-source problem, assuming that both labelled and unseen to-be-annotated images are from a single source (*e.g.* an identical website), while in practice they are generally collected from multiple sources (*e.g.* different websites). In that case, treating each source independently may suffer from the insufficiency of labelled data for model training, while merging with labelled images from other sources can bring risky biases to the source-specific model. In this paper, we propose a multi-task learning model to alleviate the multi-source image auto-annotation problem, with each task defined as performing auto-annotation for the corresponding source. Specifically, the proposed model trains annotation models for all sources in parallel with the introduction of inter-source structure regularizers and parameter constraints for sharing information and enhancing the overall performance. Experiments conducted on three different-source benchmark datasets and their combinations yield inspiring results and demonstrate that the proposed model can well utilize the shared information and relieve the risky biases.

***Index Terms*—** multi-source image annotation, multi-task learning, inter-source structure regularizers

1. INTRODUCTION

Recently the prevalence of social network and digital photography has led to the explosion of web images, necessitating effective techniques to manage and retrieve such a large-scale and rapidly-increasing image database. Image auto-annotation, which assigns proper semantic textual tags for any given image, has been revealed to be a promising approach to tackling the problem and thus attracts much attention from both academia and industry [1–12].

Previous researches on image auto-annotation can be roughly categorized into tag-view [1–3] and image-view [4–6, 8–10] models. The former treat each tag as an independent class and annotate a given image with classes it belongs to, while the latter determine the probability for each candidate tag to be associated with the given image and then adopt the ones with higher probabilities as annotations. Regarding tag-view models, E. Chang *et al.* [2] proposed a content-based soft annotation procedure by training binary classifiers for each tag, and G. Carneiro *et al.* [3] proposed to perform image auto-annotation via defining a multiclass classification problem with each tag being a class. As for image-view models, S.L. Feng *et al.* [4] proposed a generative learning approach based on multiple Bernoulli relevance model. A. Makadia *et al.* [5] proposed to perform label propagation from nearest visual neighbours. And B. Wang *et al.* [8] further integrated distance metric learning method into label propagation model for image auto-annotation.

With a survey of previous researches on image auto-annotation, we realize that most of them concentrated on the single-source problem, assuming both labelled and unseen to-be-annotated images are from a single source, *e.g.* an identical website. In real-world scenarios, however, they are generally collected from multiple different sources, *e.g.* different websites. As labelled images in each source are usually not adequate, treating each source independently will probably suffer from the insufficiency of labelled data for model training and also ignore the extra helpful information that can be exploited from other sources for performance improvement. Meanwhile, each source probably has its own distribution in the semantic space, and thus simply merging with labelled images from other sources may introduce risky biases to the source-specific model.

In this paper, we propose a multi-task learning model to tackle the problem, defining auto-annotation on each source as a task and conducting them in parallel with shared information across sources for overall performance enhancements. Multi-task learning is a learning framework aiming to improve the performance of algorithms by jointly tackling multiple tasks and utilizing their shared information, which has been successfully applied to diverse sub-fields of image processing like joint sparse feature representation [13], multi-semantic image annotation [14], *etc.* In this paper, we jointly learn linear discriminative models for all sources with inter-source structure regularizers and parameter constraints. The former regularizers enrich the correlations between low-level features and tagging vectors for a source with structure information of others, and the latter constraints force the model parameters of a shared tag in different sources to be similar.

The contributions of our work is summarized as follows: 1) We highlight the necessity of handling the multi-source image auto-annotation problem, which is more oriented to real-world scenarios. 2) We propose an effective multi-task learning model to tackle the multi-source image auto-annotation problem with inter-source structure regularizers and parameter constraints to share information across sources.

2. MULTI-SOURCE IMAGE AUTO-ANNOTATION

2.1. Problem statement

Given labelled images collected from N different sources $\left\{ \{x_j^i, y_j^i\}_{j=1}^{n_i} \right\}_{i=1}^{N}$, where n_i is the number of labelled images from the ith source and $\left\{x_j^i, y_j^i\right\}$ is the jth pair of image feature vector x_j^i and tagging vector y_j^i from the ith source, the proposed multi-source image auto-annotation model is to jointly train basic models for all sources with shared information and then annotate each unseen image with the corresponding source-specific model. Note that here the low-level image features are the same for all images, making x_j^i of an identical dimensionality for any j and i, while the vocabularies of different sources are usually diverse, meaning that the vocabulary size, *i.e.* the dimensionality of y_j^i, varies in different sources. Following [14], here we naïvely utilize the multivariate least squares regression (LSR) as a basic auto-annotation model for each source, as shown in formula (1).

$$F(M^i) = \min_{M^i} \|X^{i^T} M^i - Y^{i^T}\|_{fro}^2 \qquad (1)$$

where M^i is the parameter matrix of the ith source with each column being the regression parameters of the corresponding tag, X^i and Y^i are matrices with columns respectively being the low-level features and tagging vectors of labelled images from the source, and $\|\cdot\|_{fro}$ means the *Frobenius norm*. By minimizing the prediction error on labelled images, LSR can well model the linear correlations between image features and the value of each tag. Then the tagging vector y^i of an unseen image x^i from the same source can be predicted with $y^i = M^{i^T} x^i$.

For performance enhancements, in multi-source cases correlations and shared information across sources are supposed to be considered,which in this paper include the inter-source structure regularizers and parameter constraints.

2.2. Inter-source structure regularizers

Considering that visually similar images usually keep similarities in semantic space, their predicted tagging vectors are supposed to follow this local structure. Hence in this paper we introduce an inter-source structure regularizer for each source to utilize the structure information of others, enriching the correlations between low-level features and tagging vectors. Specifically, we construct a general k-nearest-neighbours (kNN) sparse graph as [15] consisting of images from all sources, denoted as $W_{\sum_{i=1}^N n_i \times \sum_{i=1}^N n_i}$ with $W_{s,t}$ being the similarity between image s and t if one is within the kNN of the other and zero otherwise, as shown in formula (2).

$$W_{s,t} = \begin{cases} \exp(-\|X_s - X_t\|_2^2/\sigma^2) & if\ s \sim t \\ 0 & otherwise \end{cases} \qquad (2)$$

where $s \sim t$ means s or t is within the kNN of the other, X_s and X_t are respectively the low-level feature vectors of image s and t, $\|\cdot\|_2$ means *L2 norm* and σ is the mean value of all $\|X_s - X_t\|_2$. Then for each source we take a source-specific variant of W for regularization, ignoring its intra-source image similarities, as they are supposed to be better measured with tagging vectors and thus have been implied by the basic model. Specifically, the source-specific variant for the ith source, *i.e.* W^i, is obtained by setting the similarities between images from the source in W as zero. And then its inter-source structure regularizer is formulated as follows.

$$\Phi(M^i) = \min_{M^i} \sum_{s,t} W_{s,t}^i \|X_s^T M^i - X_t^T M^i\|_2^2 \qquad (3)$$

By introducing the *Laplacian matrix* of the source-specific W^i, *i.e.* $L^i = D^i - W^i$ with D^i being a diagonal matrix and $D_{j,j}^i = \sum_k W_{j,k}^i$, the inter-source structure regularizer can be reformulated as:

$$\Phi(M^i) = \min_{M^i} Tr\left(M^{i^T} X L^i X^T M^i\right) \qquad (4)$$

where $Tr(\cdot)$ is the trace of a matrix and X is the matrix consisting of image feature vectors from all sources.

2.3. Inter-source parameter constraints

The introduction of inter-source parameter constraints is attributed to the observation that different sources generally share tags and ideally the model parameters of the predictor functions in different sources of a shared tag should be similar. Specifically, we utilize a matrix $\hat{W}_{m\times m}$ to denote the consistency between tags from all sources with $\hat{W}_{s,t} = 1$ if tag s and t are the same though from different sources and 0 otherwise, m being the sum of vocabulary sizes of all sources. Then the inter-source parameter constraints are formulated as:

$$\Psi(M) = \min_{M} \sum_{s,t} \hat{W}_{s,t} \|M_{\cdot,s} - M_{\cdot,t}\|_2^2 \qquad (5)$$

where $M = \left[M^1, \ldots, M^i, \ldots, M^N\right]$ is the concatenation of parameter matrices of different sources, and $M_{\cdot,s}$ is the sth column vector of M, *i.e.* the model parameters of tag s.

Similarly, by introducing the *Laplacian matrix* $\hat{L}$ of the tag consistency matrix $\hat{W}$, the inter-source parameter constraints can be reformulated as follows.

$$\Psi(M) = \min_{M} \; Tr\left(M\hat{L}M^T\right) \qquad (6)$$

2.4. Multi-task learning model

With a basic model for each source and the introduction of inter-source structure regularizers and parameter constraints, the proposed multi-source image auto-annotation model can be formulated as follows.

$$\begin{aligned} \mathcal{L} &= \min\nolimits_M \; F(M) + \alpha\Phi(M) + \beta\Psi(M) \\ &= \min\nolimits_M \; \sum_{i=1}^{N} C_0^i \|X^{i^T} M^i - Y^{i^T}\|_{fro}^2 + \\ &\alpha \sum_{i=1}^{N} C_1^i Tr\left(M^{i^T} X L^i X^T M^i\right) + \beta Tr\left(M\hat{L}M^T\right) \end{aligned} \qquad (7)$$

where $M = \left[M^1, \ldots, M^i, \ldots, M^N\right]$ is the concatenation of parameter matrices of different sources, α, β are factors for balancing basic models and regularizations, and C_0^i, C_1^i are source-specific constants for normalization. The introduction of C_0^i and C_1^i is to relieve biases of the optimization process, making all sources equally treated, which in our experiments are respectively set to be inversely proportional to the size of the tagging matrix and the number of structure constraints for each source, subject to $\sum_i C_0^i = 1$ and $\sum_i C_1^i = 1$.

The objective function, *i.e.* formula (7), can be demonstrated to be convex, meaning that a global optimal M^* can be found. As the number of variables can be large, the well-known Quasi Newton methods with L-BFGS updating formulas can be utilized for optimization.

3. EXPERIMENTS

3.1. Experimental settings

To evaluate the proposed model, we conduct extensive experiments on several multi-source scenarios with three well-known different-source benchmark datasets, *i.e.* Corel5k, ESPGame and IAPRTC-12, of which some statistics are shown in Table 1. We randomly take 50% of each dataset to be the training set and the rest as the test set. And the open-source Lire project [16] is utilized for image feature extraction[1], including color and textual, global and local features. Then PCA is applied for dimensionality reduction, resulting in a 844-D merged feature vector for any image, with each dimension further linearly normalized into $[0, 1]$.

The proposed multi-source image auto-annotation model, denoted as MS, is compared with the following baselines: 1) LSR-I: independent LSR on each source, 2) LSR-M: LSR with images merged from all other sources, 3) SVM-I: one-vs-all linear SVM for each tag on a source, 4) SVM-M: one-vs-all linear SVM for each tag with images merged from all other sources. To measure the performance, we follow [14] and adopt the widely-used AUC (area under ROC curve) as the metric for annotation accuracy evaluation on each tag and Mean AUC (MAUC) over tags for performance evaluation on each source. Parameter tuning for each algorithm is conducted with grid search in proper predefined parameter ranges and evaluated on randomly sampled validate sets from the training sets of all sources, *e.g.* α and β are both tuned in $\{2^{-4}, 2^{-3}, \ldots, 2^3, 2^4\}$. And we empirically utilize 20 nearest neighbours of each image to build the kNN-sparse graph.

	Corel5k	ESPGame	IAPRTC-12
Tag Nr.	256	268	291
Img. Nr.	4,999	20,770	19,627
Tags per Img.	3.4 / 5	4.7 / 15	5.7 / 19
Img. per Tag	33.2 / 542	182.2 / 2,551	192.4 / 2,783
Test Set	2,499	10,385	9,813

Table 1. Statistics of Corel5k, ESPGame and IAPRTC-12. Counts of tags and images are given as “mean / maximum”.

3.2. Experimental results

As shown in Table 2, we perform all algorithms on different combinations of the three benchmark datasets, *i.e.* different multi-source auto-annotation scenarios, α-MS and β-MS being variants of MS with only inter-source structure regularizers or inter-source parameter constraints. Then we can draw the following conclusions. 1) In most multi-source scenarios, MS and its variants outperform both LSR-based and SVM-based baselines on each source, especially LSR-I and LSR-M with the same basic model. 2) Simply merging with labelled images from other sources can promote the performance of some sources but also degrade that of others, as validated by comparing LSR-I with LSR-M and SVM-I with SVM-M. 3) Inter-source structure regularizers seem more effective than inter-source parameter constraints, since the latter only regularize the limited shared tags while the former provide more structure information for the whole vocabulary. 4) The combination of inter-source structure regularizers and parameter constraints (*i.e.* MS) just obtains slight improvement. We attribute this to that the latter constraints indirectly reflect the structure information of other sources in a high-level semantic view, which may be partly covered by the former. Yet it is interesting to find that the introduction of inter-source parameter constraints (*i.e.* β-MS) can also well improve the performance of the independent single-source basic model (*i.e.* LSR-I) and in some cases outperform the structure regularizers (*i.e.* α-MS).

To get more inside analyses, we sort the tags of each source according to the descent order of their frequencies and evenly divide them into 5 groups. Then we analyse the MAUC of each group in different multi-source scenarios.

[1] The features are: Color Layout, JCD, Edge Histogram, RGB Color Histogram and SURF with Bag-of-Words model.

Multi-Source Scenarios		LSR-I	LSR-M	SVM-I	SVM-M	α-MS	β-MS	MS
[Corel5k, ESPGame]	Corel5k	70.3	76.6	**80.4**	79.6	79.3	74.4	79.2 $\uparrow$
	ESPGame	71.5	72.0	73.6	73.6	74.0	72.2	**74.2** $\uparrow$
[Corel5k, IAPRTC-12]	Corel5k	70.3	80.1	80.4	80.5	81.5	74.6	**81.6** $\uparrow$
	IAPRTC-12	76.3	75.9	74.8	74.6	**76.8**	76.6	**76.8** $\uparrow$
[ESPGame, IAPRTC-12]	ESPGame	71.5	75.0	73.6	74.3	**76.2**	72.9	**76.2** $\uparrow$
	IAPRTC-12	76.3	75.5	74.8	74.3	76.5	**76.7**	76.5 $\uparrow$
[Corel5k, ESPGame, IAPRTC-12]	Corel5k	70.3	79.5	80.4	80.1	**81.5**	76.9	**81.5** $\uparrow$
	ESPGame	71.5	74.8	73.6	74.2	**76.4**	73.5	**76.4** $\uparrow$
	IAPRTC-12	76.3	75.4	74.8	74.2	76.5	**76.8**	76.5 $\uparrow$

Table 2. The MAUCs (%) of the proposed MS and baselines on the benchmark datasets in different multi-source auto-annotation scenarios, with "$\uparrow$" meaning MS obtains higher MAUC than LSR-I and LSR-M in the multi-source scenario.

		LSR-I	LSR-M	α-MS	β-MS	MS
Corel5k	Group 1	78.4	83.1	84.7	82.0	**84.9**
	Group 2	73.7	78.1	81.2	78.2	**81.6**
	Group 3	72.2	79.1	81.7	76.5	**81.8**
	Group 4	65.7	74.4	77.4	70.4	**77.5**
	Group 5	61.6	68.3	**71.6**	65.0	70.4
ESPGame	Group 1	76.7	77.0	**78.6**	77.5	**78.6**
	Group 2	73.3	73.8	**76.2**	74.3	**76.2**
	Group 3	70.6	71.0	**73.5**	71.5	**73.5**
	Group 4	70.7	71.3	73.2	71.4	**73.4**
	Group 5	65.8	66.5	68.3	66.3	**68.9**

Table 3. The MAUCs (%) of different tag groups of Corel5k and ESPGame in the multi-source auto-annotation scenario [Corel5k, ESPGame], with Group 1 containing the most frequent tags and Group 5 containing the most infrequent tags.

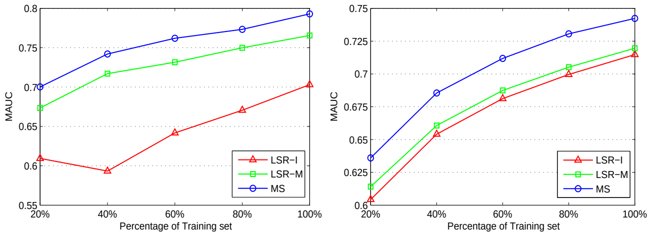

Fig. 1. MAUCs of LSR-I (red), LSR-M (green) and the proposed MS (blue) on Corel5k (left) and ESPGame (right) in multi-source scenario [Corel5k, ESPGame] with percentage of training set on both datasets varying from 20% to 100%.

As LSR is the basic model for each source, here we only consider LSR-based baselines and the proposed model. Table 3 presents the MAUC of each tag group of Corel5k and ESPGame in the multi-source scenario [Corel5k, ESPGame]. From the table, we can further see that both inter-source structure regularizers and parameter constraints can well promote most tag groups, and as expected the inter-source structure regularizers provide more promotion for tags with lower frequencies, considering that the percentage of tags with frequency over the mean value is around 25% on either dataset. Similar results are obtained in other multi-source scenarios, while they are not presented due to the space limit.

Furthermore, we randomly take different percentages of the training set of each source, *i.e.* from 20% to 100%, and evaluate the proposed model with other LSR-based baselines on the same test sets. Fig. 1 illustrates the MAUCs of different algorithms on Corel5k and ESPGame in the multi-source scenario [Corel5k, ESPGame] with different percentages of training set, from which we can conclude that the MAUCs of all algorithms mostly increase with the percentage of training set and the proposed MS maintains a clear advantage. Similar results are also obtained in other multi-source scenarios.

4. CONCLUSIONS

In this paper, we highlight the necessity of tackling multi-source image auto-annotation problem and further propose a multi-task learning model where an individual task is defined as learning a linear discriminative model for each source, with inter-source structure regularizers and parameter constraints introduced for sharing information across sources. The proposed model is evaluated with extensive experiments in several multi-source scenarios and proves its superiority to treating each source independently or simply merging with labelled data from other sources using the same basic model.

5. ACKNOWLEDGMENTS

This research was supported by the National Basic Research Project of China (Grant No. 2011CB70700) and the National Natural Science Foundation of China (Grant No. 61271394, 61005045).

6. REFERENCES

[1] J. Li and J. Wang, “Automatic linguistic indexing of pictures by a statistical modeling approach,” *IEEE Trans. PAMI*, 2003.

[2] E. Chang, K. Goh, G. Sychay, and G. Wu, “Cbsa: content-based soft annotation for multimodal image retrieval using bayes point machines,” *IEEE Trans. Circuits Syst. Video Technol.*, 2003.

[3] G. Carneiro, A.B. Chan, P.J. Moreno, and N. Vasconcelos, “Supervised learning of semantic classes for image annotation and retrieval,” *IEEE Trans. PAMI*, 2007.

[4] S.L. Feng, R. Manmatha, and V. Lavrenko, “Multiple bernoulli relevance models for image and video annotation,” in *CVPR ’04*.

[5] A. Makadia, V. Pavlovic, and S. Kumar, “A new baseline for image annotation,” in *ECCV ’08*.

[6] Z. Li, Z. Shi, Z. Li, and Z. Shi, “Modeling latent aspects for automatic image annotation,” in *ICIP ’09*.

[7] H. Wang and J. Hu, “Multi-label image annotation via maximum consistency,” in *ICIP ’10*.

[8] B. Wang, Y. Shen, and Y. Liu, “Integrating distance metric learning into label propagation model for multi-label image annotation,” in *ICIP ’11*.

[9] M. Guillaumin, T. Mensink, J. Verbeek, and C. Schmid, “Tagprop: Discriminative metric learning in nearest neighbor models for image auto-annotation,” in *ICCV ’09*.

[10] Y. Verma and C. V. Jawahar, “Image annotation using metric learning in semantic neighbourhoods,” in *ECCV ’12*.

[11] B. Bao, T. Li, and S. Yan, “Hidden-concept driven multilabel image annotation and label ranking,” *IEEE Trans. Multimedia*, 2012.

[12] Y. Yang, F. Wu, F. Nie, H.T. Shen, Y. Zhuang, and A.G. Hauptmann, “Web and personal image annotation by mining label correlation with relaxed visual graph embedding,” *IEEE Trans. Image Process.*, 2012.

[13] X. Yuan and S. Yan, “Visual classification with multitask joint sparse representation,” in *CVPR ’10*.

[14] X. Chen, X. Yuan, S. Yan, J. Tang, Y. Rui, and T. Chua, “Towards multi-semantic image annotation with graph regularized exclusive group lasso,” in *ACM MM ’11*.

[15] J. Tang, R. Hong, S. Yan, T. Chua, G. Qi, and R. Jain, “Image annotation by knn-sparse graph-based label propagation over noisily tagged web images,” *ACM Trans. Intell. Syst. Technol.*, 2011.

[16] M. Lux and S. A. Chatzichristofis, “Lire: lucene image retrieval: an extensible java cbir library,” in *ACM MM ’08*.