Renmin University of China at ImageCLEF 2014 Scalable Concept Image Annotation

Xirong Li, Xixi He, Gang Yang, Qin Jin, Jieping Xu

Multimedia Computing Lab, Renmin University of China
Key Lab of DEKE, Renmin University of China
No. 59 Zhongguancun Street, Beijing 100872, China
xirong@ruc.edu.cn

Abstract. In this paper we describe our image annotation system participated in the ImageCLEF 2014 scalable concept image annotation task. The system is fully SVM based. Per concept we learn an ensemble of fast intersection kernel SVMs from three sources of training data, all obtained with manual annotation for free. The focus of our experiments this year is to answer the question of how many tags we should use to annotate a novel image. To that end, we introduce adaptive tag selection. In contrast to the common top-k strategy which selects a fixed number of top ranked tags to annotate an unlabeled image, our method estimates the value of k with respect to the image. Given the same concept rankings, the top-5 strategy obtains MF-sample of 0.206, while adaptive tag selection reaches MF-sample of 0.311.

Keywords: Scalable image annotation, SVM, adaptive tag selection

1 Introduction

For automated image annotation, how to adaptively determine a proper number of tags for a specific image is an open problem. The top-k strategy, which simply selects the top k ranked tags per image, is probably the most popular solution. A number of systems in ImageCLEF 2013 [1] have used this strategy, including our 2013 system [2]. The fact that the number of relevant tags vary over images makes the top-k strategy suboptimal. Hence, our focus this year is on studying and evaluating strategies for adaptive tag selection.

2 The RUC 2014 System

The main components of our 2014 system, namely visual features, training data, and image annotation models, are described as follows.

2.1 Visual features

For each image, we extract bag of visual words (BoW) using the color descriptor software [3]. A precomputed codebook of size 4,000 is used to quantize densely sampled SIFT descriptors. We further consider 1x1+1x3 spatial pyramids, resulting in a BoW feature of 16,000 dimensions per image.

2.2 Training data

In addition to the 250K web images from ImageCLEF 2013, we leverage two additional two sets, both of which are acquired with manual annotation for free. The first set consists of one million images with user-clicked count, released by MSR Bing [4]. The other set consists of four million user-tagged images from Flickr. Notice that the data collecting process of the two extra sets were independent of the ImageCLEF dev/test concept lists. So the use of the extra data does not affect the scalability of our system.

2.3 Image annotation models

Different from our 2013 system [2] which combines kNN and SVM models, the 2014 edition is fully SVM based. For each dev/test tag ω, we learn an ensemble of two-class SVM classifiers from the three training sets separately.

Positive example selection. As the training sets come from different sources with different (noisy) annotation information, we describe how to select positive training examples for ω from the individual sets.

For the 250K web images, as they were collected from three web image search engines, namely Google, Yahoo, and Bing, each image x can be described by a triplet $< q, r, s >$, where q represent a query tag, r is the rank of x in the search results of q returned by an specific search engine s. Because a given image might be retrieved by different queries or by the same query but with different search engines, it can be associated with multiple triplets, denoted as $< q_i, r_i, s_i >$, $i = 1, \ldots, l$, where l is the number of triplets. To estimate the relevance of x with respect to ω, we propose to compute a search engine based score as

$$relevance_{search}(x, \omega) = \sum_{i=1}^{l} \delta(q_i, \omega) \frac{w(s_i)}{\sqrt{r_i}}, \tag{1}$$

where $\delta(q_i, \omega)$ returns 1 if q_i and ω are the same, and 0 otherwise. The variable $w(s_i)$ indicates the weight of a specific search engine, which we empirically set to be 1, 0.5, and 0.5 for Google, Yahoo, and Bing, respectively.

For the user-clicked set, each image is associated with a textual query and the accumulated count of user click. A larger click count indicates that the image is more likely to be relevant to the query [4]. We thus match ω with queries and use the corresponding click count as the relevance score.

For the Flickr set, we use a semantic-based relevance measurement as depicted in [5], which computes tagwise similarity between ω and user tags of an image.

Given the concept ω, we sort images in descending order by their relevance scores w.r.t ω, and preserve the top 1,000 ranked images as positive training examples.

SVM training. As the training data is overwhelmed by negative examples, we learn SVM classifiers by the Negative Bootstrap algorithm [6]. Different from

sampling negative examples at random, Negative Bootstrap iteratively selects negative examples which are most misclassified by present classifiers, and thus most relevant to improve classification. Per iteration, the algorithm randomly samples 10x1,000=10,000 examples to form a candidate set. An ensemble of classifiers obtained in the previous iterations are used to classify each candidate example. The top 1,000 most misclassified examples are selected and used together with the 1,000 positives to train a new classifier. For the consideration of efficiency, we use fast intersection kernel SVMs (fikSVM). For each of the three sets, we conduct Negative Bootstrap with 10 iterations, producing in total 3x10 fikSVMs per concept. These fikSVMs are further compressed into a single model such that the prediction time complexity is independent of the ensemble size.

As we focus on adaptive tag selection, we do not submit runs to compare the effectiveness of the three training sets. Nevertheless, our preliminary observation from the development set is that models trained on the three sets are complementary to each other to some extent. We therefore combine the models in a linear late fusion manner. As shown in previous studies [2, 7], weights optimized by coordinate ascent are consistently better than averaging. So we continue this good practice, and learn the fusion weights by coordinate ascent on the development set.

Adaptive tag selection. For the 107 dev concepts, we have access to a ground truth set of 1,000 images provided by ImageCLEF 2014 [8]. We find that finding a cutoff threshold that maximizes the F-concept measure per concept is a good strategy. However, this strategy is inapplicable to novel concepts that have no ground truth data available. In the 2014 task, there are 100 novel concepts, as listed in Table 1. We devise a method that selects the top k ranked tags to annotate an unlabeled image, whilst the value of k is adaptively determined with respect to the test image. The method is based on the following hypothesis. We assume that the dev concept vocabulary and the test concept vocabulary are independent, such that for a given image the ratio of its relevant tags covered in the dev vocabulary equals to the ratio of the relevant tags covered in the test vocabulary. In that regard, we can estimate the number of test concepts according to the number of dev concepts that have been selected. We will detail the method somewhere else [9].

3 Evaluation

3.1 Submitted Runs

This year we submitted eight runs:

- RUC_01: Adaptive tag selection, with the scores of test concepts updated with respect to the scores of the selected dev concepts and their cutoff threshold (Adaptive-I).
- RUC_02: Adaptive tag selection, with the scores of test concepts updated with respect to the scores of the selected dev concepts (Adaptive-II).

Table 1. The test concept vocabulary defined in the 2014 task. It consists of 107 dev concepts for which we have access to a ground truth set of 1,000 images, and 100 novel concepts which have no ground truth available.

The 107 dev concepts:
aerial airplane baby beach bicycle bird boat book bottle bridge building bus car cartoon castle cat chair child church cityscape closeup cloud cloudless coast countryside daytime desert diagram dog drink drum elder embroidery female fire firework fish flower fog food footwear forest furniture garden grass guitar harbor hat helicopter highway horse indoor instrument lake lightning logo male monument moon motorcycle mountain newspaper nighttime outdoor overcast painting park person phone plant portrait poster protest rain rainbow reflection river road sand sculpture sea shadow sign silhouette sky smoke snow soil space spectacle sport sun sunset table teenager toy traffic train tree tricycle truck underwater unpaved vehicle violin wagon water
The 100 novel concepts:
antelope apple arthropod asparagus avocado banana bear berry blood branch bread broccoli buffalo butterfly camel canidae captive carrot cauliflower cervidae cheese cheetah chimpanzee corn crocodile cucumber donkey egg eggplant elephant equidae felidae flamingo fox fried fruit galaxy giraffe gorilla grape hippopotamus human hunting kangaroo knife koala leaf leopard lettuce lion mammal marsupial meat monkey mud mushroom nebula onion orange ostrich pan pasta pear penguin pig pineapple pinniped pool potato pumpkin rabbit raccoon reptile rhino rice rifle roasted rock rodent sausage soup spider spoon squirrel strawberry submarine tiger tomato trunk tuber turtle vegetable walrus warthog watermelon wild wolf yam zebra zoo

- RUC_03: Adaptive tag selection, with the scores of test concepts updated with respect to the rank-normalized scores of the selected dev concepts (Adaptive-III).
- RUC_04: Selecting the top 5 ranked tags as final annotations.
- RUC_05: The same as RUC_01 except that the output of SVM classifiers learned from the individual data sources have been converted to probabilistic output before fusion.
- RUC_06: The same as RUC_02 except that the output of SVM classifiers learned from the individual data sources have been converted to probabilistic output before fusion.
- RUC_07: The same as RUC_03 except that the output of SVM classifiers learned from the individual data sources have been converted to probabilistic output before fusion.
- RUC_08: The same as RUC_04 except that the output of SVM classifiers learned from the individual data sources have been converted to probabilistic output before fusion.

Table 2. Performance of our eight submitted runs. Adaptive tag selection clearly outperform the top-k strategy ($k = 5$).

RunId	**Strategy**	MF-samples	MF-concepts	MAP-samples
RUC_01	Adaptive-I	28.0	24.1	**30.2**
RUC_02	Adaptive-II	27.8	24.1	**30.2**
RUC_03	Adaptive-II	27.9	24.0	**30.2**
RUC_04	Top-k	21.9	21.9	**30.2**
RUC_05	Adaptive-I	**31.1**	25.0	27.5
RUC_06	Adaptive-II	29.0	25.2	27.5
RUC_07	Adaptive-III	29.3	**25.3**	27.5
RUC_08	Top-k	20.6	21.5	27.5

3.2 Results

The performance scores of the eight runs are summarized in Table 2. Compared to the top-5 strategy (RUC_04 and RUC_08), all the other runs are better. The results clearly show that adaptive tag selection outperforms the top-k strategy.

As the only difference between RUC_01, RUC_02, RUC_03, and RUC_04 is on how to determine the value of k, which does not change concept ranking, the MAP-samples scores are the same for the four runs. Similarly, RUC_05, RUC_06, RUC_07, and RUC_08 have the same MAP-samples scores. Comparing the two groups of runs, we find that score normalization before fusion is helpful for improving MF-samples and MF-concepts, but causes performance drop in MAP-samples.

4 Conclusions

This paper documents our experiments in the ImageCLEF 2014 Scalable Concept Image Annotation, a testbed for developing a scalable image annotation system with manual annotation for free. A novel component of our 2014 system is adaptive tag selection that determines the number of tags to be selected with respect to a given test image. Adaptive tag selection is better than selecting a fixed number of tags for image annotation.

Acknowledgments. This research was supported by the National Science Foundation of China (No. 61303184), the Fundamental Research Funds for the Central Universities and the Research Funds of Renmin University of China (No. 14XNLQ01), the Specialized Research Fund for the Doctoral Program of Higher Education (No. 20130004120006), and the Scientific Research Foundation for the Returned Overseas Chinese Scholars, State Education Ministry. The authors are grateful to the ImageCLEF coordinators for the benchmark organization efforts [8, 10].

References

1. Villegas, M., Paredes, R., Thomee, B.: Overview of the imageclef 2013 scalable concept image annotation subtask. In: CLEF 2013 working notes. (2013)
2. Li, X., Liao, S., Liu, B., Yang, G., Jin, Q., Xu, J., Du, X.: Renmin university of china at imageclef 2013 scalable concept image annotation. In: ImageCLEF working notes. (2013)
3. van de Sande, K., Gevers, T., Snoek, C.: Evaluating color descriptors for object and scene recognition. IEEE Trans. Pattern Anal. Mach. Intell. **32** (2010) 1582–1596
4. Hua, X.S., Yang, L., Wang, J., Wang, J., Ye, M., Wang, K., Rui, Y., Li, J.: Clickage: Towards bridging semantic and intent gaps via mining click logs of search engines. In: ACM MM. (2013)
5. Li, X., Snoek, C., Worring, M., Smeulders, A.: Harvesting social images for bi-concept search. IEEE Transactions on Multimedia **14**(4) (Aug. 2012) 1091–1104
6. Li, X., Snoek, C., Worring, M., Koelma, D., Smeulders, A.: Bootstrapping visual categorization with relevant negatives. IEEE Transactions on Multimedia **15**(4) (Jun. 2013) 933–945
7. Li, X., Snoek, C., Worring, M., Smeulders, A.: Fusing concept detection and geo context for visual search. In: ICMR. (2012)
8. Villegas, M., Paredes, R.: Overview of the ImageCLEF 2014 Scalable Concept Image Annotation Task. In: CLEF 2014 Evaluation Labs and Workshop, Online Working Notes. (2014)
9. He, X., Li, X., Yang, G., Jin, Q., Xu, J.: Adaptive top-k tag selection for image annotation. (2014) submitted.
10. Caputo, B., Müller, H., Martinez-Gomez, J., Villegas, M., Acar, B., Patricia, N., Marvasti, N., Üsküdarlı, S., Paredes, R., Cazorla, M., Garcia-Varea, I., Morell, V.: ImageCLEF 2014: Overview and analysis of the results. In: CLEF proceedings. Lecture Notes in Computer Science. Springer Berlin Heidelberg (2014)