Renmin University of China at ImageCLEF 2013 Scalable Concept Image Annotation

Xirong Li[1,2], Shuai Liao[1], Binbin Liu[2], Gang Yang[2]
Qin Jin[2], Jieping Xu[2], Xiaoyong Du[1]

[1]Key Lab of Data Engineering and Knowledge Engineering, MOE
[2]Multimedia Computing Lab, School of Information, Renmin University of China
Zhongguancun Street 59, Beijing 100872, China
xirong@ruc.edu.cn

Abstract. In this paper we describe our image annotation system participated in the ImageCLEF 2013 scalable concept image annotation task. The system leverages multiple base classifiers, including single-feature and multi-feature kNN classifiers and histogram intersection kernel SVMs, all of which are learned from the provided 250K web images and provided features with no extra manual verification. These base classifiers are combined into a stacked model, with the combination weights optimized to maximize the geometric mean of F-samples, F-concepts, and AP-samples metrics on the provided development set. By varying the configuration of the system, we submitted five runs. Evaluation results show that for all of our runs, model stacking with optimized weights performs best. Our system can annotate diverse Internet images purely based on the visual content, at the following accuracy level: F-samples of 0.290, F-concepts of 0.304, and AP-samples of 0.380. What is more, a system-to-system comparison reveals that our system and the best submission this year are complementary with respect to the best annotated concepts, suggesting the potential for future improvement.

Keywords: Image annotation, learning from web, stacked model

1 Introduction

Annotating *unlabeled* images by computers is crucial for organizing and retrieving the ever-growing amounts of images on personal devices and the Internet. Due to the semantic gap, i.e., the lack of coincidence between visual features extracted from the visual data and a user's interpretation on the same data, image auto-annotation is challenging.

In the context of annotating Internet images, the semantic gap becomes even bigger, as a specific concept exhibits significant diversity in its visual appearance. The imagery of a concept does not limit to realistic photographs, but can also be artificial correspondences such as posters, drawings, and cartoons, as demonstrated in Fig. 1. To annotate the uncontrolled visual content with a large set of concepts, a promising line of research is to learn from web data which contains

airplane cloud daytime poster sky vehicle

bird cartoon

bird flower painting plant

Fig. 1. Internet images and ground truth tags from the development set of the Scalable Concept Image Annotation 2013 task.

many images but with unreliable annotations [1–6]. In these works, k Nearest Neighbors (kNN) and Support Vector Machines (SVMs) are two popular classifiers, as have been separately used in [1–3] and [4–6]. Given the difficulty of the Scalable Concept Image Annotation 2013 task [7], we believe that a single classifier is inadequate. In that regard, we develop an image annotation system that combines a number of base classifiers into a stacked model. In our 2013 experiments we submitted five runs, with the purpose of verifying the effectiveness of model stacking.

The remainder of the paper is organized as follows. We first describe our image annotation system in Section 2. Then we detail our experiments in Section 3, with conclusions given in Section 4.

2 The RUC Image Annotation System

Our image annotation system consists of multiple base classifiers, which are then combined into a stacked model to make final predictions. The base classifiers are learned from the 250K Internet images provided by the 2013 task, while the stacked model is optimized using the development set of 1,000 ground-truthed images. A conceptual diagram of the system is given in Fig. 2.

Next, we describe the stacked model in Section 2.1, followed by the base classifiers in Section 2.2.

2.1 Stacked Image Annotation Model

For the ease of consistent description, let x be a specific image. For a given concept ω, let $g(x, \omega)$ be an image annotation model which produces a relevance score of ω with respect to x. A set of t models are denoted by $\{g_1(x, \omega), \ldots, g_t(x, \omega)\}$.

We define our stacked model as a convex combination of the t models:

$$g_\Lambda(x, \omega) = \sum_{i=1}^{t} \lambda_i \cdot g_i(x, \omega), \tag{1}$$

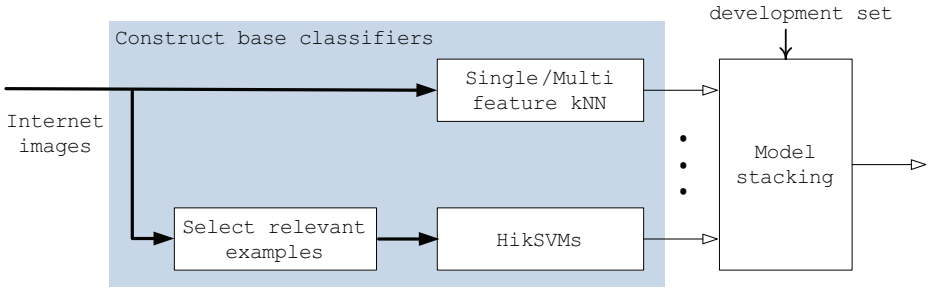

Fig. 2. A conceptual diagram of the RUC image annotation system participated in the Scalable Concept Image Annotation 2013 task.

where λ_i is the nonnegative weight of the i-th classifier, with $\sum_{i=1}^{t} \lambda_i = 1$. Notice that $g_\Lambda(x, \omega)$ is to indicate that the stacked model is parameterized by $\Lambda = \{\lambda_i\}$.

We look for the setting of $\{\lambda_i\}$ that maximizes the image annotation performance. The 2013 task specifies three performance metrics [7]: F-measure for images, F-measure for concepts, and Average Precison for images, as given in the Appendix. To jointly maximize the three metrics, we take their *geometric mean* as a combined metric. Weights of Eq. (1) optimized with respect to the combined metric are found by a coordinate ascent algorithm.

2.2 Base Classifiers

We choose k Nearest Neighbors (kNN) and Support Vector Machines (SVMs) as two types of base classifiers, for their good performance.

Base classifier I: kNN. Given a test image x, we define the kNN classifier as

$$g_{knn}(x, \omega) = \sum_{i=1}^{k} rel(x_i, \omega) \cdot w_i, \tag{2}$$

where $rel(x_i, \omega)$ denotes the relevance score between the i-th neighbor and the concept, and w_i is the neighbor weight. In this work, we instantiate $rel(x_i, \omega)$ using the provided textual feature, which is derived from the web page of x_i [7]. For w_i, we choose a Bayesian implementation [3], computing w_i as $k^{-1} \sum_{j=i}^{k} j^{-1}$ to give more importance to closer neighbors.

Base Classifier II: SVMs. We choose the histogram intersection kernel SVMs (hikSVMs) which is known to effective for bag of visual words features of medium size [8], such as the 5,000-dim features provided by the task. More importantly, the decision function of a hikSVMs can be efficiently computed by a few linear interpolations on a set of precomputed points.

To obtain relevant positive training examples for a given concept from the 250K Internet images, we utilize both the textual feature and the query log generated when collecting the images from web image search engines. We describe

the connection of an Internet image x to a web image search engine s by a triplet $< q, r, s >$, where q represents a query keyword, r is the rank of x in the search results of q returned by s. An image can be associated with multiple triplets. To determine the positiveness of the image x with respect to the given concept ω, we propose to compute a search engine based score as

$$\text{positiveness}(x, \omega) = \sum_{i=1}^{l} sim(q_i, \omega) \frac{w(s_i)}{\sqrt{r_i}}, \tag{3}$$

where l is the number of triplets associated with the image, and $sim(q_i, \omega)$ is a tag-wise similarity measure, computed on the base of tag co-occurrence in 1.2 million Flickr images [4]. 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. The positiveness score in Eq. 3 is further combined with the given textual feature. For each concept, we sort images labeled with the concept in descending order by their positiveness scores, and select the top 500 ranked images as positive training examples.

For the acquisition of negative training examples, we consider the following two approaches. The first approach is to sample negative examples at random. Given a specific concept and the 500 selected positive examples, we randomly sample a set of 500 negative examples, to make the training data perfectly balanced. We repeat the random sampling 10 times, yielding a set of 10 hikSVMs. The second approach is Negative Bootstrap [9]. Different from random sampling, this approach iteratively selects negative examples which are most misclassified by present classifiers, and thus most relevant to improve classification. Per iteration, the approach randomly samples 5,000 examples to form a candidate set. An ensemble of classifiers obtained in the previous iterations are used to classify each candidate example. The top 500 most misclassified examples are selected and used together with the 500 positives to train a new hikSVMs. We conduct Negative Bootstrap with 10 iterations, producing 10 hikSVMs for each concept. For efficient classification, we leverage a model compression technique [10] to compress the ensemble into a single classifier such that the prediction time complexity is independent of the ensemble size.

3 Experiments

3.1 Submitted Runs

This year we submitted five runs, which are listed as follows. For all runs, we empirically preserve for each image the top six ranked concepts as its final annotation.

Run: RUC Crane. This run uses a kNN classifier with early fused multiple features (Multi-Feat-kNN). We use all the 7 provided features, i.e., getlf (256-D), colorhist (576-D), gist (480-D), sift (5000-D), csift (5000-D), rgbsift (5000-D),

and opponentsift (5000-D). For each feature, we compute the l_1 distance. Distance values of individual features are zero-score normalized, and then averaged to obtain $k = 256$ nearest neighbors.

Run: RUC Snake. This run combines multiple single feature kNN (Single-Feat-kNN) and SVMs classifiers with uniform weights. In particular, five features (colorhist, gist, csift, rgbsift, and opponentsift) are used separately for Single-Feat-kNN, again with the l_1 distance and $k = 256$. For SVMs, we use the three variants of sift, i.e., csift, rgbsift, and opponentsift. In total, this run employs 5+3+3=11 base classifiers.

Run: RUC Monkey. This run uses the same 11 base classifiers as have been used in the *Snake* run, but with optimized weights. Consequently, the effectiveness of the stacked model can be verified by comparing the *Monkey* and the *Snake*.

Run: RUC Mantis. This run combines the Multi-Feat-kNN, multiple Single-Feat-kNN, and 6 variants of SVMs with uniform weights. So the run employs 12 base classifiers in total.

Run: RUC Tiger. This run uses the same 12 base classifiers as have been used in the *Mantis* run, but with optimized weights. The *Tiger* is our primary run.

3.2 Results

The performance scores of the five runs are summarized in Table 1. For all performance metrics, the *Tiger* run, which combines all the base classifiers using the coordinate ascent optimized weights, is the best. On the test set, this run reaches an F-samples score of 0.290, an F-concepts score of 0.304, and an AP-samples score of 0.380. The *Monkey* run using optimized weights is better than the *Snake* run using uniform weights. Moreover, by comparing the performance on the development set and on the test set, we find that the system generalizes well to unseen images and concepts. These results justify the importance of model stacking and weights optimization for image annotation.

Table 1. Settings and performance of our submitted runs. The symbol + indicates that a specific base classifier is used in a specific run. Our primary run *Tiger*, which combines all base classifiers with optimized weights, performs best.

	Based classifiers				**F-samples**		**F-concepts**		**AP-samples**	
Run	Multi-Feat-kNN	Single-Feat-kNN	SVMs	**Weights**	dev	test	dev	test	dev	test
Crane	+	-	-	N.A.	28.8	25.4	26.6	23.9	36.1	32.4
Snake	-	+	+	uniform	28.8	26.5	30.8	28.5	38.2	35.5
Monkey	-	+	+	optimized	31.0	28.3	32.7	29.6	40.5	37.6
Mantis	+	+	+	uniform	29.8	27.8	31.4	29.2	39.4	36.9
Tiger	+	+	+	optimized	**31.6**	**29.0**	**33.4**	**30.4**	**41.2**	**38.0**

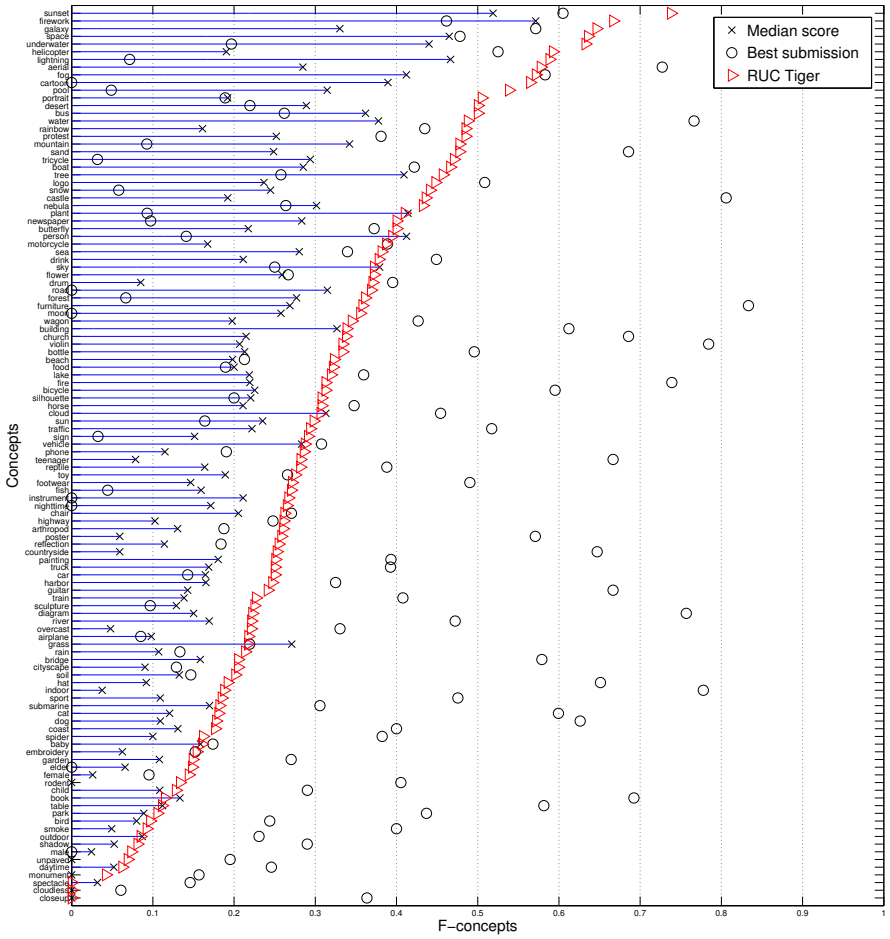

Fig. 3. A system-to-system comparison. The concepts have been sorted in terms of their F-concepts scores. The median score of each concept is the median of the F-concepts scores of all the 58 submissions. The *RUC Tiger* run is clearly above the average level. What we find more interesting is that our system and the best submission seem complementary, showing the potential for future improvement.

Fig. 3 shows a system-to-system comparison, measured by F-concepts. The RUC image annotation system is clearly above the average level. Moreover, when compared with the best submission of this year, we find that the two systems are complementary, as their best annotated concepts differ.

4 Discussions and Conclusions

This paper documents our experiments in the ImageCLEF 2013 Scalable Concept Image Annotation, a testbed for developing image annotation systems using generic web data. We have built such a system.

Our system annotates images purely based on the visual content. It combines multiple base classifiers, i.e., variants of kNN and SVMs, into a stacked model. In all of our five submitted runs, the stacked model with optimized weights performs best.

Through a system-to-system comparison, we find that our system and the best submission this year is complementary in the sense of the best annotated concepts. Given that we use relatively simple base classifiers, we consider this finding interesting, as it suggests the potential of future improvement.

Acknowledgments. This research was supported by the Basic Research funds in Renmin University of China from the central government (13XNLF05). The authors are grateful to the ImageCLEF coordinators for the benchmark organization efforts.

Appendix

A1. Performance Measures

F-samples. Given a test image x with its relevant tag set R_x and a predicted tag set P_x, its F-samples score is computed as

$$\text{F-samples}(x) = \frac{2 * \text{precision}(x) * \text{recall}(x)}{\text{precision}(x) + \text{recall}(x)}, \tag{4}$$

where precision(x) is $|R_x \cap P_x|/|P_x|$, and recall(x) is $|R_x \cap P_x|/|R_x|$.

F-concepts. Given a test concept ω with its relevant image set R_ω and a set of images P_ω labeled with ω by the annotation system, its F-concepts score is computed as

$$\text{F-concepts}(\omega) = \frac{2 * \text{precision}(\omega) * \text{recall}(\omega)}{\text{precision}(\omega) + \text{recall}(\omega)}, \tag{5}$$

where precision(ω) is $|R_\omega \cap P_\omega|/|P_\omega|$, and recall($\omega$) is $|R_\omega \cap P_\omega|/|R_\omega|$.

AP-samples. Given a test image x with m tags sorted in descending order by relevance scores, its AP-samples score is computed as

$$\text{AP-samples}(x) = \frac{1}{|R_x|} \sum_{i=1}^{m} \frac{r_i}{i} \delta(i), \tag{6}$$

where r_i is the number of relevant tags among the top i tags, and $\delta(i)$ is 1 if the i-th tag is in R_x, 0 otherwise.

References

1. Wang, X.J., Zhang, L., Li, X., Ma, W.Y.: Annotating images by mining image search results. IEEE Transactions on Pattern Analysis and Machine Intelligence **30**(11) (Nov. 2008) 1919–1932
2. Li, X., Snoek, C., Worring, M.: Annotating images by harnessing worldwide user-tagged photos. In: ICASSP. (2009)
3. Villegas, M., Paredes, R.: A k-nn approach for scalable image annotation using general web data. In: BigVision. (2012)
4. 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
5. Zhu, S., Jiang, Y.G., Ngo, C.W.: Sampling and ontologically pooling web images for visual concept learning. IEEE Transactions on Multimedia **14**(4) (Aug. 2012) 1068–1078
6. Kordumova, S., Li, X., Snoek, C.: Evaluating sources and strategies for learning video concepts from social media. In: CBMI. (2013)
7. Villegas, M., Paredes, R., Thomee, B.: Overview of the imageclef 2013 scalable concept image annotation subtask. In: CLEF 2013 working notes. (2013)
8. Maji, S., Berg, A., Malik, J.: Classification using intersection kernel support vector machines is efficient. In: CVPR. (2008)
9. Li, X., Snoek, C., Worring, M., Smeulders, A.: Social negative bootstrapping for visual categorization. In: ICMR. (2011)
10. 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