MLIA at ImageCLFE 2014 Scalable Concept Image Annotation Challenge

Xing Xu, Atsushi Shimada, and Rin-ichiro Taniguchi

Department of Advanced Information Technology, Kyushu University, Japan
{xing, atsushi}@limu.ait.kyushu-u.ac.jp, {rin}@ait.kyushu-u.ac.jp

Abstract. In this paper, we propose a large-scale image annotation system for the ImageCLEF 2014 Scalable Concept Image Annotation task. The annotation task, of this year, concentrated on developing annotation algorithms that rely only on data obtained automatically from the web. Since the sophisticated SVM based annotation techniques had been widely applied in the task last year (ImageCLEF 2013), for the task this year, we also adopt the SVM based annotation techniques and put our effort mainly on obtaining more accurate concepts assignment for training images. More specifically, we proposed a two-fold scheme to assign concepts to unlabeled training images: (1) A traditional process which stems the extracted web data of each training image from textual aspect, and make concepts assignment based on the appearance of each concept. (2) An additional process which leverages the deep convolutional network toolbox Overfeat to predict labels (in ImageNet nouns) for each training image from visual aspect, then the predicted tags are mapped to concepts in ImageCLEF based on WordNet synonyms and hyponyms with semantic relations. Finally, the allocated concepts for each training image are generated based on a fusion step of the two-fold concepts assignment processes. Experimental results show that the proposed concepts assignment scheme is efficient to improve the assignment results of traditional textual processing and to allocate reasonable concepts for training images. Consequently, with an efficient SVMs solver based on Stochastic Gradient Descent, our annotation systems achieves competitive performance in the annotation task.

Keywords: imageclef, image annotation, social web data

1 Introduction

In this year ImageCLEF 2014 [1], we participated the Scalable Concept Image Annotation challenge[1] [10] which aimed at developing more scalable image annotation system. The goal of this challenge is to develop annotation systems that for training only rely on unsupervised web data and other automatically obtainable resources. In contrast to traditional image annotation evaluations with labeled training data, this challenge requires work in more front, such as handling

[1] http://www.imageclef.org/2014/annotation

the noisy data, textual processing and multilabel annotations and scalability to unobserved labels.

Since this year is the third edition of the annotation challenge, regarding the methodology of annotation system, we can make several observations from the overview reports [9] [11] of previous editions:

- The best performing system, TPT [6], only used provided visual features, which indicated that the visual features provided by the organizers is sufficient enough and the other features extracted by several teams might be complementary.
- The top 3 teams (TPT, MIL [4], and UNIMORE [2]) all utilized SVMs based algorithms to learn separate classifiers for each concept, which was verified to be superior to the K nearest neighbor (KNN) based annotation techniques used by other groups, such as RUC [5], MICC [8].
- The textual processing and concepts assignment for training images were significant, since they directly affected the learning accuracy of concept classifiers.

The major difference of the challenge this year compared with previous editions is the proportions of "scaled" concepts. In the challenge last year, there are total 116 concepts (95 concepts for development set and 21 more for test set), the proportions of "scaled" concepts are $\frac{21}{116} \approx 0.181$. On the contrast, in this year, there are total 207 concepts (107 concepts for development set and 100 more for test set), the proportions of "scaled" concepts are $\frac{100}{207} \approx 0.483$. Thus it implies the significance of annotation system to be scalable and to generalize well to the new concepts.

To develop a robust and scalable annotation system, we believe that one of the intrinsic issues is to assign more appropriate concepts to training images. Once we have collected more accurate (positive/negative) samples for each concept, it is possible to improve the performance of concepts' classifiers. Thus for the contest, we mainly focus on the issue of accurate concepts assignment for training images. Besides the traditional textual information processing such as stopwords removal and stemming, which have been widely applied in previous editions. We also leverage the recent popular convolutional neural networks (CNN) [7] to allocate tags (1K WordNet nouns) for each training images from visual aspect. As the CNN based method utilizes the deep neural network to improve classification task, we can rely on the tags predicted by Overfeat and map the tags to concepts of ImageCLFE vocabulary. Then a late fusion approach is used to decide the final concepts assignment for each training image. Finally, we train a linear SVM classifier for each concept (similar in development and test set) with the visual features provided by the organizers. To tackle the high dimensional large volumes of training data, we adopt the online learning strategy of stochastic gradient descent (SGD). We finally obtain competitive annotation performances in terms of mAP-samples, MF-concepts and MF-samples measures and are ranked the 4th place among all 11 groups on overall measure.

The rest of the paper is organized as follows. Section 2 demonstrates the architecture of proposed annotation system and we mainly discuss our concepts

assignment scheme for training images. In Section 3, we describe our experimental setups and report the evaluation results obtained on both the development and the test sets. And Section 4 includes conclusion and some future direction of our works.

2 Proposed annotation system

The proposed annotation system is depicted in Figure 1. To assign more appropriate concepts for training images, we conduct a 2-fold scheme which explicitly leverages the provided textural information semantically (Section 2.1) and the training images visually (Section 2.2). Based on the reliable labeled training images, we further learn SVMs based concept classifiers using standard visual features provided by the organizers. To tackle the high dimensional features and large volumes of data, we use online learning method combined with SGD algorithm. Then we use the learnt stable concept classifiers for concept prediction of images in development and test sets. In the following subsections, we would like to depict the detailed procedure of each module of the diagram in Figure 1.

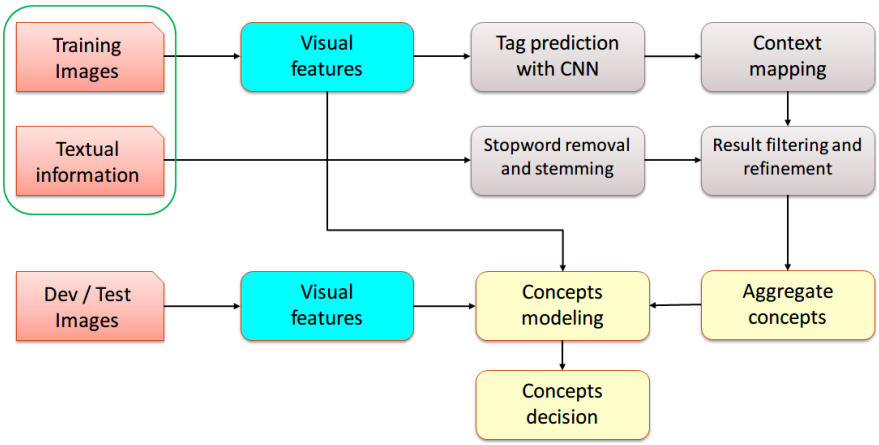

Fig. 1. Overview of proposed annotation system architecture

2.1 Text Processing Approach

The organizers of ImageCLFE 2014 provided several kinds of textural features of training images. Following the traditional text processing approach utilized last year, to efficiently process the textual features, we applied multiple filtering on the textural features. Regarding the modules of "Stopword removal and stemming" in Figure 1, the detailed processing procedures are:

– "Stopword removal and stemming" is performed on the "scofeats" files, where stopwords, misspelled words, words from different languages other than English, the titles of the original web pages are extracted and parsed.
– We then matched the semantic relations of the remaining words with the list of concepts in development set based on WordNet 3.0[2]. We extend the list of concepts with their synonyms, and examine whether current word matches with concept or its synonyms.
– The Lucene [3] stemmer is adopted if the word does not exactly match with the list of concepts.

The output of the "Result filtering and refinement" produces a candidate set of concepts for each of the training image. Indeed, the processing approach in this subsection could be considered as a baseline as it gives many false negative and false positive concepts to training images. Therefore, besides the textual features of training images, it is reasonable to further consider the visuality of training images. For example, for a training image describing "airplane", and its textural features (web page, title, etc) contain words of "airplane pilot hats", simply applying the text processing approach would result in concepts "airplane", "person", and "hat" to be assigned to the training image. However, if it is possible to estimate the content of image visually in advance, then the unrelated concepts "person", "hat" could be rejected to the training image. Thus, in the next subsection, we would like to introduce a context mapping method to predict tags for training images in advance.

2.2 Context mapping using CNN

To estimate the content of training images visually, we take advantages of a recently proposed toolbox Overfeat[3], which is an image recognizer and feature extractor built around a deep convolutional neural network (CNN). We consider this powerful toolbox for two reasons: (1) It achieved competitive classification results on ImageNet 2013 contest[4]. (2) OverFeat convolutional net was trained on WordNet 1K nouns, which is consistent to the concept list of ImageCLFE. Thus it is rational to predict tags for training images based on the Overfeat and mapping the tags to ImageCLFE using a built context mapping rule. Regarding the modules "Tag prediction with CNN" and "Context mapping" in Figure 1, the detailed processing procedures are:

– For a given training image, we directly use the Overfeat toolbox to predict tags for it.
– For each of the tag predicted from Overfeat, we calculate its semantic similarity to the concept list of development set, and mapping it to the most similar concept.

[2] http://wordnet.princeton.edu/
[3] http://cilvr.nyu.edu/doku.php?id=software:overfeat:start
[4] http://www.image-net.org/challenges/LSVRC/2013/results.php

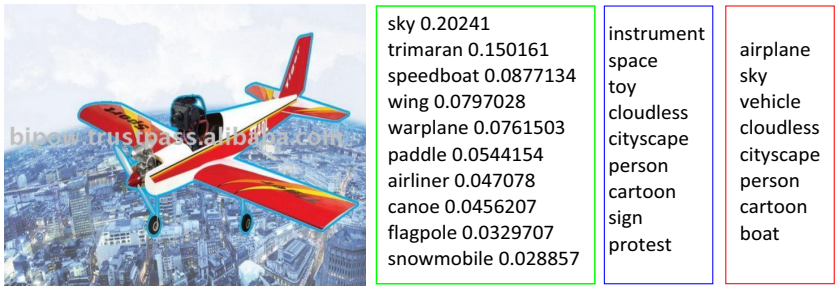

Fig. 2. Example of context mapping using CNN on a training image.

In Figure 2, we give an example of the context mapping using CNN. The tags in blue rectangle are obtained from the previous "text processing" stage. The tags (with confidence scores) in green rectangle are tags predicted from Overfeat. For the context mapping procedure (in practice, we use the *path_similarity* measure in NLTK toolbox as the semantic measure), we can get a candidate concept set {*sky*, *airplane*, *vehicle*, *boat*} based on the tags in green rectangle .

For the "Result filtering and refinement" module in Figure 1, it fuses the candidate concept set from both textual processing approach and context mapping with CNN. Since there are much more number of concepts produced by textual processing approach than context mapping with CNN, however, the concept set from textual processing approach is more coarse. Thus for the fusion strategy, we relied more on the concept set from context mapping with CNN and preserved the concepts with high similarity scores in concept set from textual processing approach. In Figure 2, the concepts in red rectangle are the final assigned concepts to the training image, which are considered to be semantically related to the training image.

3 Experimental results

3.1 Visual features

Similar as the best result of TPT [6] in ImageCLFE 2013 annotation task, we use the visual features provided by the organizer including GIST, Color Histogram, SIFT, C-SIFT, RGB-SIFT and OPPONENT-SIFT. For all SIFT-based descriptors, a bag-of-words (BoW) representation is provided. An early fusion is made by concatenating all the features provided (global color histogram, getlf, CSIFT, GIST, opponent SIFT, RGB-SIFT, SIFT) resulting in a 21,312 dimension space. Global features GIST and Color Histogram are normalized using L2 norm, and SIFT-based features are normalized using L1 norm.

3.2 Evaluation measures

For the performance measures used to evaluate the runs, there are three standard measures: mean F-measure for the samples (MF-samples), mean F-measure for

the concepts (MF-concepts) and the mean average precision for the samples (MAP-samples). The MF is computed analyzing both the samples (MF-samples) and the concepts (MF-concepts), whereas the MAP is computed analyzing the samples.

3.3 Training SVM classifiers for concepts

Following the SVM based annotation techniques which had achieved best annotation performance last year [2] [6], again we trained "one-versus-all" SVM classifier for each concept. The popular SVM solvers, such as SVMlight, LibSVM, they are not feasible for training large volumes of data with high dimension, since these batch methods need to pre-load entire training data into memory, to compute gradient in each iteration. Thus it is difficult to directly utilize these SVM solvers. According to the configuration of our machine (an Intel Core i7 2600 CPU (3.4 GHz) and 16 GB RAM), we take into account a better solution by the stochastic gradient descent (SGD) algorithm which is more efficient for training SVM classifiers with large-scale data. Different from the batch method, in the SGD algorithm, training sample is fed one by one to calculate the gradients and update rules of model parameters. Although the SGD algorithm might need more iteration loops to reach convergence, it requires much less memory cost which is more appropriate for large-scale training samples and online learning manner.

According to the advices in [2], we randomize the training data and load the data in chunks which fit in memory, then train the different classifiers on further randomizations of chunks, so that different epochs will get the chunks data with different ordering which leads the learnt classifiers to be stable. We repeat this training process on training set for 5 times to train SVM classifier for each concept of development set and cross validate the F-measure on development set. Then we select the parameters of best performance on development set to further learn classifiers for concepts of test set. To predict concepts for images in development and test sets, we use the trained concepts's classifiers and obtain decision scores for each concept by thresholding the confidence score at zero.

3.4 Inside analysis of annotation results

We first discuss the proposed 2-fold concept assignments to training images, and evaluate its influence of learning accuracy of concept classifiers. We first conduct experiments on the development set and then extend the 2-fold scheme to test set. Here we consider three settings: (1) "Single-Fold A": the single fold scheme of traditional textual information process ("Stopword removal and stemming" module in Figure 1). (2) "Single-Fold B": the single fold scheme of CNN based tag prediction process ("Tag prediction with CNN" and "Context mapping" modules), (3) "Two-Folds": the fusion process of both "Single-Fold A" and "Single-Fold B". We limited the maximum number of concepts assigned to each training image to be 4. Then we use the learned SVMs classifiers from the labeled training data to predict concepts for images in development set (the top

5 ranked concepts are considered to be the final predicted concepts). Table 1 shows the annotation performance of three settings, and one of the baselines provided by the organizers is also included for comparison. It can be observed that three settings consistently improve the performance of baseline. In particular, the tags predicted by Overfeat is considerably accurate for training images. "Single-Fold B" outperforms the "Single-Fold A" setting of traditional textual information scheme, which implies the tags is highly coherent with the concepts in ImageCLFE. Moreover, when fusing the two settings to formulate proposed "Two-Folds" setting, the result is further improved on all three measures.

Table 1. Annotation results on development set: three settings of textual information processing scheme of concept assignments for training images.

Run	MF-sample	MF-concept	MAP-sample
Baseline (SIFT)	0.1342	0.2261	0.2254
Single-Fold A	0.218	0.203	0.3321
Single-Fold B	0.2693	0.2445	0.3622
Two-Folds	0.3105	0.3224	0.3781

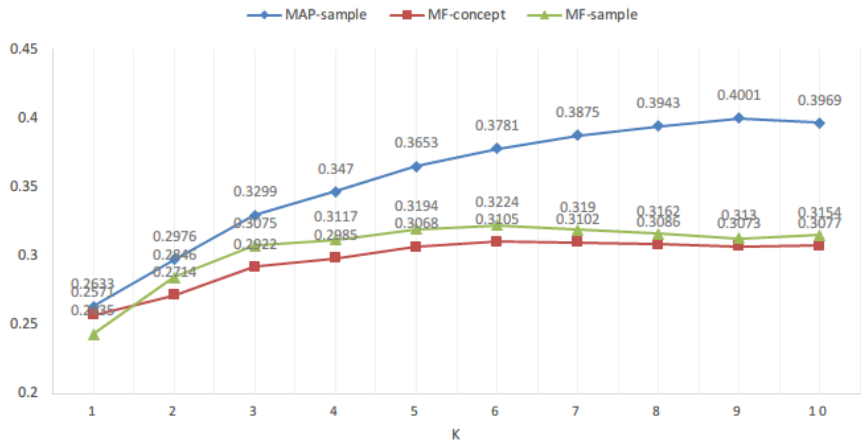

Fig. 3. Annotation performance on development set with varying K.

Then we evaluate the effect of "Result filtering and refinement" module. Since in the experiment settings above, we restrict the number (denoted by K) of assigned concepts to each training image as $K = 4$. And it is reasonable that the value of K could influence the learning accuracy of concept classifiers, as it directly determines the quality of training samples for each concept. Thus,

we further vary the value of K (ranges from 1 to 10), and explore the optimal K for concept assignments for training images. The annotation performance on development set with varying K is shown in Figure 3. It can be observed from Figure 3 that: (1) The peaks of both MF-concept and MF-sample are reached when $K = 6$, and peak of MAP-sample reaches the peak when $K = 9$. (2) The MAP-sample is more sensitive to K since the number of ground truth concepts for each image in development set ranges from 1 to 11 (with average 3.52). Based on these observations, finally we choose $K \in [6, 7, 8, 9, 10]$ for our latter submit runs of test set.

For the test set, we submitted ten runs[5]. Here we would like to present our best 5 runs with baselines provided by organizers and the best runs from the other groups. We can learn from the overall results in Table 2 that: (1) All our submitted runs are beyond the best baseline result for the test set according to all measures. Looking into the overall participants results list, our best runs are at position 6, 3 and 5 order by the MF-sample, MF-concept and MAP-sample respectively for the test set, and position 4 for the overall performance. It means that our best runs are competitive compared with other results.

Table 2. Annotation results of our best 5 runs on the test set, compared best runs of baselines and other groups.

Run	MF-sample	MF-concept	MAP-sample
Baseline (oppsift)	16.7	9.8	20.2
kdevir_09	37.7	54.7	36.8
MIL_03	27.5	34.7	36.9
MindLab_01	25.8	30.7	37
DISA-MU_04	29.7	19.1	34.3
RUC_05	31.1	25	27.5
IPL_09	18.4	15.8	23.4
IMC-FU_01	16.3	12.5	25.1
INAOE_05	5.3	10.3	9.6
NII_01	13	2.3	14.7
FINKI_01	7.2	4.7	6.9
MLIA_09	24.8	33.2	27.8
MLIA_10	24.8	33.2	27.9
MLIA_08	24.6	33.3	27.4
MLIA_07	24.4	33.5	26.9
MLIA_06	24.1	33.6	26.3

However, there is still a considerable gap between our best runs and the top-ranked runs from KDEVIR group. Although currently we are not able to explore the details of their proposed annotation technique, there are still space to improve our annotation system itself from the following aspects: (1) In our current

[5] `http://www.imageclef.org/2014/annotation/results`

system, we directly utilized the Overfeat toolbox for tag prediction of training images, a more reasonable choice is that we can generate CNN visual features and directly use these visual features to learn concept classifiers. Indeed, several teams such as MIL and MindLab used the CNN visual features. (2) Currently, the "Context mapping" module only considered mapping the tags from Overfeat to ImageCLEF with its synonymous/hyponyms in WordNet, and the similarity measure from NLTK toolbox might not be precise to map the correct results. An optional choice is modeling the context based similarity measure of tags depending on the Flickr image metadata, which is more efficient to capture the semantic associations from the practical circumstance. (3) Our concept modeling (SVMs based concept classifiers learning) is not elaborately optimized and tuned, because of the limitations of hardware configurations and consumption of resources. Our system capability should be improved if we could overcome these limitations.

4 Conclusion

In this paper, we presented our annotation system developed to participate at ImageCLEF 2014 for the Scalable Concept Image Annotation task. Our proposal focus on improving the accuracy of concept assignments for training images. We proposed a 2-fold concept assignments scheme which explicitly leverages the provided textural information semantically (Section 2.1) and the training images visually. To learn concept classifiers, we adopted the sophisticated SVM based model, and took the SGD algorithm to deal with large scale settings of this task. Experimental results show that our proposal on both visual and textual information processing are necessary to build a competitive system. Moreover, we also considered potential future directions to further improve current system.

References

1. 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)
2. Grana, C., Serra, G., Manfredi, M., Cucchiara, R., Martoglia, R., Mandreoli, F.: Unimore at imageclef 2013: Scalable concept image annotation. In: CLEF 2013 Evaluation Labs andWorkshop, OnlineWorking Notes (2013)
3. Hatcher, E., Gospodnetic, O., McCandless, M.: Lucene in action. Second Edition
4. Hidaka, M., Gunji, N., Harada, T.: Mil at imageclef 2013: Scalable system for image annotation. In: CLEF 2013 Evaluation Labs andWorkshop, OnlineWorking Notes (2013)
5. 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: CLEF 2013 Evaluation Labs and Workshop, Online Working Notes (2013)
6. Sahbi, H.: Telecom paristech at imageclef 2013 scalable concept image annotation task: Winning annotations with context dependent svms. In: CLEF 2013 Evaluation Labs and Workshop, Online Working Notes (2013)

7. Sermanet, P., Eigen, D., Zhang, X., Mathieu, M., Fergus, R., LeCun, Y.: Overfeat: Integrated recognition, localization and detection using convolutional networks. CoRR abs/1312.6229 (2013)
8. Uricchio, T., Bertini, M., B., L., Bimbo, A.: Micc at imageclef 2013 image annotation subtask. In: CLEF 2013 Evaluation Labs andWorkshop, Online Working Notes (2013)
9. Villegas, M., Paredes, R.: Overview of the imageclef 2012 scalable concept image annotation subtask. In: CLEF 2012 Evaluation Labs and Workshop, Online Working Notes (2012)
10. 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)
11. Villegas, M., Paredes, R., Thomee, B.: Overview of the imageclef 2013 scalable concept image annotation subtask. In: CLEF 2013 Evaluation Labs and Workshop, Online Working Notes. pp. 1–19 (2013)