KDEVIR at ImageCLEF 2015 Scalable Image Annotation, Localization, and Sentence Generation task: Ontology based Multi-label Image Annotation

Md Zia Ullah[1] and Masaki Aono[†]

Department of Computer Science and Engineering,
Toyohashi University of Technology,
1-1 Hibarigaoka, Tempaku-Cho, Toyohashi, 441-8580, Aichi, Japan,
arif@kde.cs.tut.ac.jp[1], aono@tut.jp[†]

Abstract. In this paper, we describe our participation in the ImageCLEF 2015 Scalable Concept Image Annotation task. In this participation, we propose an approach of image annotation by using ontology at several steps of supervised learning with noisy unlabeled data. In this regard, we construct tree-like ontology for each annotating concept of images using WordNet and Wikipedia. The constructed ontologies are exploited throughout the proposed framework including several phases of training and testing of one-vs-all SVM classifiers. Several classifiers are trained on local or global visual features separately and results are ensemble using the classifiers' probability scores. The result turns out that our system achieves an average performance in this task.

Keywords: Image Annotation, Classification, Feature-wise learning, Ontology

1 Introduction

Due to the explosive growth of digital technologies, collections of images are increasing tremendously in every moment. The ever growing size of the image collections has evolved the necessity of image retrieval (IR) systems; however, the task of IR from a large volume of images is formidable since binary stream data is often hard to decode, and we have very limited semantic contextual information about the image content.

To enable the user for searching images using semantic meaning, automatically annotating images with some concepts or keywords using machine learning is a popular technique. During last two decades, there are a large number of researches being lunched using state-of-the-art machine learning techniques [1–4] (e.g. SVMs, Logistic Regression). In such efforts, most often each image is assumed to have only one class label. However, this is not necessarily true for real world applications, as an image might be associated with multiple semantic tags. Therefore, it is a practical and important problem to accurately assign

multiple labels to one image. To alleviate above problem i.e. to annotate each image with multiple labels, a number of research have been carried out; among them adopting probabilistic tools such as the Bayesian methods is popular [5–7]. More review can be found in [8, 9]. However, accuracy of such approach depends on expensive human labeled training data.

Fortunately, some initiatives have been taken to reduce the reliability on manually labeled image data [10–13] by using cheaply gathered web data. Although the "Semantic gaps" between low-level visual features and high-level semantics still remain and accuracy is not improved remarkably.

In order to reduce the dependencies of human-labeled image data, ImageCLEF [14] has been organizing the image annotation task for the last several years, where training data is a large collection of Web images without ground truth labels. Despite the proposed methods in this task shown encouraging performance on a large scale dataset, unfortunately none of them utilizes the semantic relations among annotating concepts.

In this paper, we describe the participation of KDEVIR at ImageCLEF 2015 Scalable Image Annotation, Localization, and Sentence Generation task [15], where, we have focused on image annotation subtask. In this regard, we have proposed an approach, ontology based learning that exploits both textual and visual features of images during training and testing. The evaluation results reveal the effectiveness of proposed framework.

The rest of the paper is organized as follows: **Section 2** describes the proposed framework. **Section 3** describes our submitted runs to this task as well as comparison results with other participants' runs. Finally, concluded remarks and some future directions of our work are described in **Section 4**.

2 Proposed Framework

In this section, we describe our method for annotating images with a list of semantic concepts. We divide our method into four steps: 1) Constructing Ontology, 2) Pre-processing of Training Data, 3) Training Classifier, and 4) Predicting Annotations. An overview of our proposed framework is depicted in Fig. 1.

2.1 Constructing Ontology

Ontologies are the structural frameworks for organizing information about the world or some part of it. In computer science and information science, ontology is defined as an explicit, formal specification of a shared conceptualization [16, 17] and it formally represents knowledge as a set of concepts within a domain, and the relationships between those concepts. To utilize these relationships in image annotation, we construct ontology for each concept of a predefined list of concept used to annotate images.

In real world, an image might contain multiple objects (aka concepts) in a single frame, where concepts are inter-related and maintain a natural way of being co-appearance. We use these hypotheses to construct ontologies for

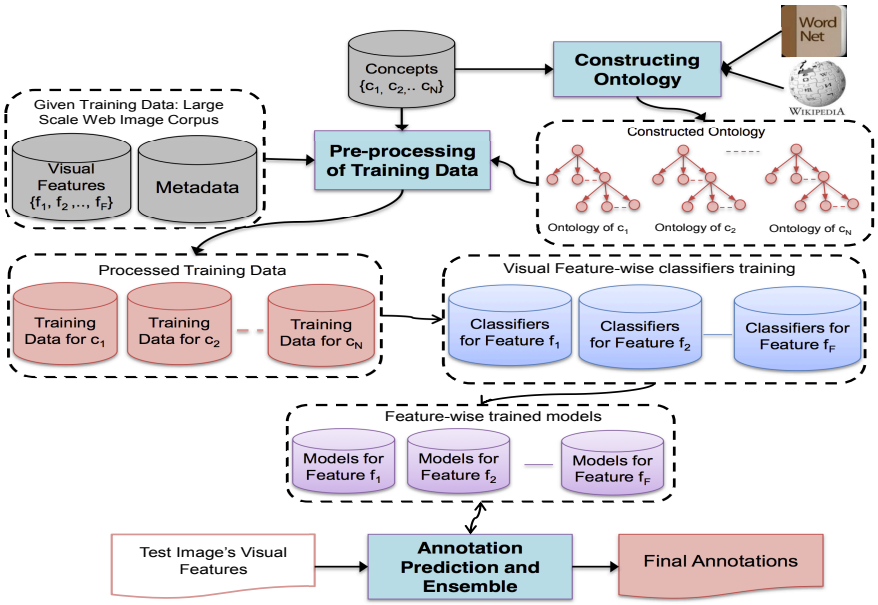

Fig. 1: Proposed Framework

concepts [18]. In this regard, we utilize WordNet [19] and Wikipedia as primary sources of knowledge.

Let C be a set of concepts. We construct a tree-like ontology [20] for each concept $c_c \in C$. In order to build ontologies, first of all, we select some types of relations including: 1) taxonomical R_t, 2) functional R_f, and 3) weak hierarchical, R_{wh}. The relations are extracted empirically according to our observations on WordNet and Wikipedia articles. For each type of relations, we extract a set of relationship property as listed below:
R_t= {*"inHypernymPathOf"*, *"subClassOf"*, *"isA"*}
R_f={*"habitat"*, *"inhabit"*, *"liveIn"*, *"foundOn"*, *"foundIn"*, *"locateAt"*, *"nativeTo"*, *"liveOn"*, *"feedOn"*}
R_{wh}={*"kindOf"*, *"typeOf"*, *"representationOf"*, *"methodOf"*, *"appearedAt"*, *"appearedIn"*, *"ableToProduce"*}

Finally, we apply some *"if-then"* type inference rules to add an edge from a parent-concept to a child-concept by leveraging the above relations.

2.2 Pre-processing of Training Data

Given a list of concepts, we select the potential images for each concept from the noisy training images by exploiting their metadata (details about metadata are given in [15]) and pre-constructed concept ontologies. In this regards, first of all, we detect the nouns and adjectives from metadata using WordNet followed

by singularizing with Pling Stemmer[1]. Secondly, detected terms from metadata: Web text (scofeat), keywords, and URLs are weighted by BM25 [21], mean reciprocal rank (MRR), and a constant weight, $\vartheta \in (0, 1)$ respectively, which is followed by detecting concepts from the weighted sources on appearance basis. Thus, we have three lists of possible weighted concepts from three different sources of metadata for each image.

We take the inverted index of image-wise weighted concepts, thus generate the concept-wise weighted images. To aggregate the images for a concept from three sources, we normalize the weight of images using Max-Min normalization technique, and linearly combine the BM25, MRR, and constant ϑ to generate the final weight of images. From the resultant aggregated list of images, top-m images are primarily selected for each concept.

Finally, in order to increase the recall, we merge the primarily selected training images of each concept with its parent concepts of highest semantic confident (i.e. parents connected by $r_t \in R_t$) by leveraging our concept ontologies. Thus, we enhance training images per-concept as well as number of annotated concepts per-image.

2.3 Training Classifier

Image annotation is a multi-class multi-label classification problem; current state-of-the-art classifiers are not able to solve this problem in their usual format. Towards this problem, we propose a technique of using ontologies during different phases of learning a classifier. In this regard, we choose Support Vector Machines (SVMs) as a classifier for its robustness of generalization. We subdivide the whole problem into several sub-problems according to the number of concepts, i.e. train SVMs for each concept separately, since using a large dataset at a time is not rational in terms of memory and time.

Another problem is that, along with the different parameters, the classification accuracy of SVMs depends on the positive and negative examples which are used to train the classifier. It is obvious that if classifiers are trained with wrong examples, the prediction will be wrong. However, selecting appropriate training example is formidable without any semantic clues. In this regard, for a concept, we take positive examples from its image-list which is generated in the preprocessing stage and the negative examples from all other concepts' image-lists those are not semantically related to the current concept. To handle this issue, we use our pre-constructed concept ontologies.

For each local or global visual feature, we train one-vs-all SVM for all concepts. With positive and negative examples, we train $|F|$ probabilistic one-vs-all SVM models for each concept, where F is a set of visual feature types including CNN, GIST, Color Histograms, SIFT, C-SIFT, RGB-SIFT, and OPPONENT-SIFT. We use LIBSVM [22] to learn the SVM models. As kernel, instead of using the default choice of Linear kernel or Gaussian kernel, since image classification

[1] http://www.mpi-inf.mpg.de/yago-naga/javatools/index.html

is a nonlinear problem and distribution of image data is unknown, we choose histogram intersection kernel (HIK) [23]. The HIK is defined as:

$$k_{HI}(h^{(a)}, h^{(b)}) = \sum_{q=1}^{l} min(h_q^{(a)}, h_q^{(b)}) \quad (1)$$

where $h^{(a)}$ and $h^{(b)}$ are two normalized histograms of l bins; in context of image data, two feature vectors of l dimensions.

2.4 Predicting Annotations

The trained models for all concepts generated based on each visual features in the previous subsection are used to predict annotations. Given a test image, if a model of particular concept responds positively, the image is considered as voted by current model i.e. the corresponding concept is primarily selected for annotation. At the same time, the tracks of predicted probability and vote are kept. This process is repeated for all learned models for all concepts. The concept-wise predicted probabilities and votes are accumulated for all visual features. In second level selection, empirical thresholds for accumulated probabilities and votes are used to select more relevant annotations. Finally, we take top-k weighted concepts as annotation for the test image. In ImageCLEF 2015 [15], the test dataset and train dataset are same. This makes the concept detection of test data possible by using only the textual features of train dataset. In our proposed framework, we have both textual and visual features to recognize test images. However, in experiments, we conducted some runs using only the textual features of train data to annotate the test images. These runs confirm the validity of our preprocessing of noisy training data.

3 KDEVIR Runs and Comparative Results

We submitted total five runs, which are differ from each other in terms of: use of ontology or not; number of primarily selected training images, m; and based on textual, visual features or both; number of $topK$ concepts selected for annotations. The configurations of all runs are given in Table 1, where runs are arranged according to their original name to ease the flow of description. Here, run 1, 2, and 3 are employed based on both textual and visual features. However, run 4 and 5 are constructed based on the textual features of trained data, because both train and test dataset are same.

In Table. 2, evaluation results of our submitted runs are illustrated. It reveals that "run 4" produces the best performance in terms of mean average precision (MAP), although we did not use any visual features in this run. It shows the effectiveness of our preprocessing stage of training data. However, the performance of "run 1", "run 2", and "run 3" are not satisfactory. It turns out that feature-wise learning of several visual features is not effective, although we could not afford to process all visual features including CNN and SIFT variants due to

Table 1: Configurations of our submitted runs. The run pairs **Run** {**1**, **2**, and **3** were conducted to show the effectiveness of using ontology and visual features with Histogram intersection kernel; while, the run pairs **Run** {**4** and **5**} were conducted to show the effect of data preprocessing using ontology and weighting methods.

Run	**Ontology?**	**Visual Feature**	**Kernel**	***m***	***topK***
Run 1	Yes	ColorHist, GETLF, GIST	HIK	3000	10
Run 2	Yes	ColorHist, GETLF, GIST	HIK	3000	15
Run 3	Yes	OpponentSIFT, ColorHist, GETLF, GIST	HIK	3000	15
Run 4	No	Textual Feature	No	1000	20
Run 5	Yes	Textual Feature	No	1000	20

time constrain. Either, we need more effective visual features or more efficient kernel and ensemble methods to boost up the performance of image annotation. Details about all the performance measures are given in [15].

Table 2: Evaluation results of our submitted runs in terms of MAP_0.5 overlap and MAP_0 overlap.

Run	**MAP_0.5 Overlap**	**MAP_0 Overlap**
Run 1	0.019876	0.048681
Run 2	0.021726	0.050720
Run 3	0.024631	0.055277
Run 4	**0.228856**	**0.386693**
Run 5	0.14093	0.305518

4 Conclusion

In this paper, we described the participation of KDEVIR at ImageCLEF 2015 Scalable Concept Image Annotation task, where we proposed an approach for annotating images using ontologies at several phases of supervised learning from large scale noisy training data.

The evaluation result reveals that our proposed approach achieved an average performance among all submitted runs in terms of MAP_0.5 and MAP_0 overlap measures. However, in some runs, our system performance is not satisfactory. We could not afford to process all visual features due to time constraint. In future, we will consider deep learning to detect concepts in the noisy web images.

Acknowledgement

This research was partially supported by the HORI FOUNDATION of JAPAN, Grant-in-Aid C114.

References

1. Dumont, M., Marée, R., Wehenkel, L., Geurts, P.: Fast multi-class image annotation with random windows and multiple output randomized trees. In: Proc. International Conference on Computer Vision Theory and Applications (VISAPP) Volume. Volume 2. (2009) 196–203
2. Alham, N.K., Li, M., Liu, Y.: Parallelizing multiclass support vector machines for scalable image annotation. Neural Computing and Applications **24**(2) (2014) 367–381
3. Qi, X., Han, Y.: Incorporating multiple svms for automatic image annotation. Pattern Recognition **40**(2) (2007) 728–741
4. Park, S.B., Lee, J.W., Kim, S.K.: Content-based image classification using a neural network. Pattern Recognition Letters **25**(3) (2004) 287–300
5. Rui, S., Jin, W., Chua, T.S.: A novel approach to auto image annotation based on pairwise constrained clustering and semi-naïve bayesian model. In: Multimedia Modelling Conference, 2005. MMM 2005. Proceedings of the 11th International, IEEE (2005) 322–327
6. Yang, C., Dong, M., Fotouhi, F.: Image content annotation using bayesian framework and complement components analysis. In: Image Processing, 2005. ICIP 2005. IEEE International Conference on. Volume 1., IEEE (2005) I–1193
7. Jeon, J., Lavrenko, V., Manmatha, R.: Automatic image annotation and retrieval using crossmedia relevance models. (2003)
8. Gong, Y., Jia, Y., Leung, T., Toshev, A., Ioffe, S.: Deep convolutional ranking for multilabel image annotation. arXiv preprint arXiv:1312.4894 (2013)
9. Zhang, D., Islam, M.M., Lu, G.: A review on automatic image annotation techniques. Pattern Recognition **45**(1) (2012) 346–362
10. Cai, D., He, X., Li, Z., Ma, W.Y., Wen, J.R.: Hierarchical clustering of www image search results using visual, textual and link information. In: Proceedings of the 12th annual ACM international conference on Multimedia, ACM (2004) 952–959
11. Gupta, M.R., Bengio, S., Weston, J.: Training highly multiclass classifiers. Journal of Machine Learning Research **15** (2014) 1–48
12. Weston, J., Bengio, S., Usunier, N.: Large scale image annotation: learning to rank with joint word-image embeddings. Machine learning **81**(1) (2010) 21–35
13. Wang, X.J., Zhang, L., Jing, F., Ma, W.Y.: Annosearch: Image auto-annotation by search. In: Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on. Volume 2., IEEE (2006) 1483–1490
14. Villegas, M., Müller, H., Gilbert, A., Piras, L., Wang, J., Mikolajczyk, K., de Herrera, A.G.S., Bromuri, S., Amin, M.A., Mohammed, M.K., Acar, B., Uskudarli, S., Marvasti, N.B., Aldana, J.F., del Mar Roldán García, M.: General Overview of ImageCLEF at the CLEF 2015 Labs. Lecture Notes in Computer Science. Springer International Publishing (2015)
15. Gilbert, A., Piras, L., Wang, J., Yan, F., Dellandrea, E., Gaizauskas, R., Villegas, M., Mikolajczyk, K.: Overview of the ImageCLEF 2015 Scalable Image Annotation, Localization and Sentence Generation task. In: CLEF2015 Working Notes. CEUR Workshop Proceedings, Toulouse, France, CEUR-WS.org (September 8-11 2015)

16. Gruber, T.R.: Toward principles for the design of ontologies used for knowledge sharing? International journal of human-computer studies **43**(5) (1995) 907–928
17. Studer, R., Benjamins, V.R., Fensel, D.: Knowledge engineering: principles and methods. Data & knowledge engineering **25**(1) (1998) 161–197
18. Reshma, I.A., Ullah, M.Z., Aono, M.: Kdevir at imageclef 2014 scalable concept image annotation task: Ontology based automatic image annotation
19. Miller, G.A.: Wordnet: a lexical database for english. Communications of the ACM **38**(11) (1995) 39–41
20. Wei, W., Gulla, J.A.: Sentiment learning on product reviews via sentiment ontology tree. In: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics (2010) 404–413
21. Robertson, S.E., Walker, S., Beaulieu, M., Willett, P.: Okapi at trec-7: automatic ad hoc, filtering, vlc and interactive track. Nist Special Publication SP (1999) 253–264
22. Chang, C.C., Lin, C.J.: Libsvm: a library for support vector machines. ACM Transactions on Intelligent Systems and Technology (TIST) **2**(3) (2011) 27
23. Swain, M.J., Ballard, D.H.: Color indexing. International journal of computer vision **7**(1) (1991) 11–32