MICC-UNIFI at ImageCLEF 2013 Scalable Concept Image Annotation

Tiberio Uricchio, Marco Bertini, Lamberto Ballan, and Alberto Del Bimbo

Media Integration and Communication Center (MICC)
Università degli Studi di Firenze, Italy
{name.surname}@unifi.it

Abstract. In this paper we report on MICC participation to the Scalable Concept Image Annotation subtask of the ImageCLEF Photo Annotation and Retrieval competition [13].
Our goal has been to investigate the applicability of data-driven methods that have obtained good results in the field of social image annotation and retrieval to web images. These methods have been applied typically to tasks such as tag ranking, tag suggestion and refinement. Since they do not require a training stage they can be applied in cases in which the set of annotation keywords can vary greatly over time or when the set of images to be analysed is very large.

Keywords: Image annotation, image tagging, social media.

1 Introduction

This paper describes our participation in the Scalable Concept Image Annotation subtask of the 2013 ImageCLEF competition [2]. It is a standardized benchmark for systems that automatically annotate images based on a varying vocabulary and a large corpus of web images with their corresponding web pages [13]. No annotated ground truth data is available, except for a small dataset which is used exclusively to test the system during its development. We submitted five runs using an unsupervised scalable approach based on nearest-neighbors by experimenting with several parameters.

Recently, data-driven approaches have shown to be able to deal with very large scale scenarios, and have been applied to tag ranking for social image retrieval, tag refinement for social image annotation [9, 7, 4, 12]. In order to address the problem of large-scale collections and maintaining an efficient approach, we choose to evaluate the use of such nearest-neighbor approaches also in the context of web images annotation.

Our approach, described in section 3, computes a visual distance between test images and train images and then obtains a score for several words in WordNet by performing a simple density estimation. Afterwards, a final score for each of the concepts from the requested vocabulary is obtained by evaluating several semantic similarities. Section 3 describes in more detail the various steps of the method; Section 4 reports the experimental setup used, while description of runs and results are reported in Section 5. Conclusions are drawn in Section 6.

2 Method Overview

Given a set of training images I with their respective web pages, a set of test images I_T and a vocabulary of words V the goal is to get a relevance value $r(i, w) \forall w \in V, \forall i \in I_T$ and to choose a set of final annotations to be assigned. The latter ones can be simply obtained by using a threshold on relevance values or by enumerating the first fixed N words, where N has been determined empirically. Our method is comprised of four steps:

1. Building a set of artificial tags for every image in the provided training set. This casts the problem as a tag refinement task.
2. Features extraction from training and test images and computation of an image – and corresponding web pages – neighborhood for every test image.
3. Construction of a candidate words set by image annotation, based on TagRelevance method over text features.
4. Filtering of stop-words and re-ranking of words by using several semantic metrics, defined on WordNet and Wikipedia ontologies.

The obtained scores correspond to the final relevance value assigned. Particular attention is given to the issue of scalability: our approach can scale up to utilize as many features and data as possible. Note that we are also assuming an open-world vocabulary which comprises potentially every possible word used on the web. However, as English WordNet and English Wikipedia are more mature, we consider only words which are contained in these two ontologies. This limitation can be possibly overcame by exploiting a system for automatic translation like Google Translate or Bing Translator; however, we have not used this approach in this work.

3 A Nearest Neighbor approach

The basic idea of nearest-neighbor methods is to select a set of visually similar images and then to select a set of relevant associated words based on a word transfer procedure. Images selected as visually similar must be tagged with a set of words possibly related to the content. This type of methods has also been applied to different tasks such as tag suggestion and tag ranking/relevance, applied to the context of social media [7, 4, 12]. There is no need to use an explicit training of a model as it is implicitly made by the choosing of distance and space.

In this ImageCLEF task, however, training images don't have any words (or tags) associated, instead they have one or more web pages with natural language text content. To overcame this issue, the first step is to build a training set of artificially labeled images to be used as a source of neighbors. Text, metadata and URLs from web pages are transformed in a set of tags for every image. It is not required to have perfect annotations as nearest neighbors method can make use of bigger training set, by simply using a bigger visual neighborhood sample to better estimate the specific tag distribution. A source of noise is related to the kind of relationship between the images and corresponding web pages: some

images can possibly be completely disassociated to the content described in the web page. As there's no easy way to determine if this is the case, some images can possibly be artificially tagged with completely unrelated tags. In our experiments we directly employed a set of text features provided.

3.1 Learning Tag Relevance from Visual Neighbors: Li *et al.* [7]

Li *et al.* have proposed a tag relevance measure for image retrieval based on the consideration, originally proposed in [5], that if different persons label visually similar images using the same tags, then these tags are more likely to reflect objective aspects of the visual content. Therefore it can be assumed that the more frequently the tag occurs in the set of images that form the visual neighborhood of the image to be annotated, the more relevant it might be. However, some frequently occurring tags are unlikely to be relevant to the majority of images. To account for this fact the proposed tag relevance measurement takes into account both the distribution of a tag t in the neighbor set for an image I and in the entire collection:

$$tagRelevance(t, I, K) := n_t[N_k(I, K)] - Prior(t, K) \tag{1}$$

where n_t is an operator counting the occurrences of t in the neighborhood $N_k(I, K)$ of K similar images, and $Prior(t, K)$ is the occurrence frequency of t in the entire collection. In order to reduce user bias, only one image per different user is considered when computing the visual neighborhood. As the neighborhood increases in size, it can be proved that tags selected by TagRelevance yields to an ideal image ranking [7], provided that probability to sample visually similar image is greater than sampling random images. The method has been tested for image retrieval on a proprietary Flickr dataset with 20,000 manually checked images and for image auto-annotation using a subset of 331 images. Recently it has been applied to a bigger social media dataset named NUSWIDE-240K [12], showing considerable performance for image retagging.

Differently from the original approach of [7] we weight the occurrences of t with the distance: considering the setup of the auto-annotation experiment, we estimate *tagRelevance* for each candidate tag and then rank the tags in descending order by *tagRelevance*. Given a test image I the procedure used for tag refinement is:

1. Estimation of the distribution of each tag t of I in $N_k(I, K)$.
2. Computation of *tagRelevance* of each tag t subtracting $Prior(t, K)$ from the distribution of t in $N_k(I, K)$.
3. Ranking of the tags according to their *tagRelevance* score.
4. Transfer the n highest ranking tags.

4 Experiments

The core of a working nearest neighbor approach is given by the space where images are represented and by the selection of a good distance measure. In

our experiments we use only a visual space derived directly from the visual features provided with the ImageCLEF datasets. 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. All features are singularly normalized using ℓ_2 norm. In our implementation the distance between images is computed using a Gaussian kernel:

$$d(I_i, I_k) = \frac{e^{||\mathbf{f}_i - \mathbf{f}_k||}}{\sigma^2} \tag{2}$$

where I_i is the visual neighbor in the i position, with N features $\mathbf{f}_i = (f_i^1, \ldots, f_i^N)$, and σ is set as the median value of all the distances. The size of Development and Test sets, together, is only of $3,000$ images and the training set is comprised of $250,000$ images, which constitute a total of $3,000 \times 250,000 = 750,000,000$ distances. Given the relatively low number, we directly computed all distances exhaustively. The process took about three hours on a medium spec computer. In alternative, as the training set increases, one can use an approximate technique like LSH[3][11], without losing too much precision. Assuming to measure distance in double-precision floats, a matrix of this dimension needs $7,5{*}10^8{*}8 = 6{*}10^9 \sim 6$ GB of RAM, a relatively big size for a medium spec machine. To ease working, we precomputed distances between every image (either from development or test set) to all training images, retaining only the $4,999$ nearest. This results in two distance matrices of respectively $1,000 \times 4,999$ and $2,000 \times 4,999$ for development and test set of about 115 MB.

After some initial experiments we have used the square of the distance also to weight the occurrences of a tag t in the neighborhood of an image I_k.

4.1 Text Features

We used three kind of provided features: the score features, the triplets used to get images in search engines and the training URLs to create the set of tags associated with the images. The URLs were processed to extract the words that composed them by means of regular expressions and by checking their presence in WordNet.

4.2 Semantic Augmentation

To cope with the fact that the tag refinement approach used is applied to images that are associated with a set of textual features that could be different from the set of keywords to be used for annotation, we have tested some approaches to perform simple semantic augmentation of the tags resulting from the process described in Section 3.1.

Initially we have tried to add WordNet synonyms to the list of selected tags. However this approach has resulted in very limited improvement. A second approach has provided some steady improvement when using the Test dataset and therefore has been used also in all the runs on the Development dataset: 10 tags

with the highest tag relevance score are selected, then the ImageCLEF keywords that have an overall strong semantic similarity with them are added to the list of candidate words used for annotation. This selection is performed by computing the average semantic distance between all the tags, considering the best semantic relatedness based on Wikipedia article internal links structure [10], and then selecting the ImageCLEF keywords with a lower average distance.

4.3 Fusion Methods

Nearest neighbor approaches have proven to be able to use several distances by fusing the results of more than one system [8]. Inspired by this we tried to fuse several runs where parameters were different in number of neighbors (from 50 to 4,999), text features selected (several combinations of all features) and different distances (ℓ_1, ℓ_2, χ_2). Following [6], given the result of several classifiers $X_1, X_2, \ldots, X_k \in \mathbb{R}^{I \times C}$, where I is the number of images and C is the number of categories, we employed simple fusion techniques $Y = \text{operator}(X_1, X_2, \ldots, X_k)$ without learned parameters, for completely unsupervised classifiers. We tried average, multiplication, max and min, followed or preceded by a soft-max operation. Another technique we tried is that of Borda count [1], a well-known rank aggregation algorithm. However, none of the combinations resulted in more than very limited improvement, ranging from losing $3-5\%$ to improving 0.5% at the cost of several runs of executions of all the single modalities.

5 Description of Runs and Discussion

We submitted five runs, using both Development and Test datasets. In all the runs we used all the pre-computed features to evaluate the visual neighborhood, score features were used as image tags and tag relevance was computed weighting the presence of tags using the squared visual distance between the image to be annotated and its visual neighborhood.

Only two parameters were varied during the runs: the number of tags assigned to each image and the size of the visual neighborhood.

1. Run 1: 2000 NNs, 5 tag per image.
2. Run 2: 3000 NNs, 10 tag per image.
3. Run 3: 3000 NNs, 7 tag per image.
4. Run 4: 4000 NNs, 7 tag per image.
5. Run 5: 4999 NNs, 7 tag per image.

Results in terms of F1 micro, F1 macro and MAP are reported for Development and Test datasets in Table 1 and 2, respectively. It can be observed that the larger the number of visual neighbors the better the performance. The improvement is much reduced on the Test dataset probably due to the over fitting induced by using the same set of images from the Training dataset to compute the visual neighborhood in both experiments.

The system has been completely developed in Python, without attempting to implement any particular optimization. Running all the experiments on a portable PC with 2.53 GHz Intel Core i5 processor takes about 2.5 hours.

Run	F1 micro	F1 macro	MAP
1	20.4	20.3	28.7
2	23.3	20.7	29.0
3	22.3	21.0	29.0
4	22.4	21.0	29.2
5	22.7	21.4	29.1

Table 1. Experimental results of the 5 runs on the Development dataset

Run	F1 micro	F1 macro	MAP	F1 macro unseen concepts
1	18.7	17.3	25.9	17.6
2	20.4	17.5	26.1	17.0
3	20.0	18.1	26.1	18.5
4	20.0	18.0	26.1	18.6
5	20.0	18.0	26.2	18.6

Table 2. Experimental results of the 5 runs on the Test dataset. The 5^{th} column reports results on the set of concepts that were not part of the Development set.

6 Conclusion

In this paper we have presented our system for web images annotation based on a data-driven approach that has been used for tag reranking in the context of social media. Thanks to its simplicity and the fact that it requires no training or supervision, the system can be executed on mid level PCs and can be easily applied to other datasets. The system has also just two main parameters that have to be adjusted: the number of images used to create the visual neighborhood of the images to be annotated and the number of tags to be selected for annotation.

References

1. Aslam, J.A., Montague, M.: Models for metasearch. In: Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval. pp. 276–284. ACM (2001)
2. Caputo, B., Muller, H., Thomee, B., Villegas, M., Paredes, R., Zellhofer, D., Goeau, H., Joly, A., Bonnet, P., Gomez, J.M., Varea, I.G., Cazorla, M.: ImageCLEF 2013: the vision, the data and the open challenges. In: Proc. of CLEF, LNCS (2013)
3. Gionis, A., Indyk, P., Motwani, R., et al.: Similarity search in high dimensions via hashing. In: Proceedings of the international conference on very large data bases. pp. 518–529 (1999)
4. Guillaumin, M., Mensink, T., Verbeek, J., Schmid, C.: Tagprop: Discriminative metric learning in nearest neighbor models for image auto-annotation. In: Proc. of ICCV (2009)

5. Kennedy, L.S., Slaney, M., Weinberger, K.: Reliable tags using image similarity: mining specificity and expertise from large-scale multimedia databases. In: Proc. of ACM-MM Workshop on Web-Scale Multimedia Corpus. Beijing, China (2009)
6. Kuncheva, L.I.: Combining Pattern Classifiers: Methods and Algorithms. Wiley-Interscience (2004)
7. Li, X., Snoek, C.G.M., Worring, M.: Learning social tag relevance by neighbor voting. IEEE Transactions on Multimedia 11(7), 1310–1322 (2009)
8. Li, X., Snoek, C.G.M., Worring, M.: Unsupervised multi-feature tag relevance learning for social image retrieval. In: Proc. of ACM CIVR (2010)
9. Makadia, A., Pavlovic, V., Kumar, S.: A new baseline for image annotation. In: Proc. of ECCV (2008)
10. Milne, D., Witten, I.H.: An effective, low-cost measure of semantic relatedness obtained from Wikipedia links. In: Proc. of AAAI (2008)
11. Muja, M., Lowe, D.G.: Fast approximate nearest neighbors with automatic algorithm configuration. In: International Conference on Computer Vision Theory and Application VISSAPP'09). pp. 331–340. INSTICC Press (2009)
12. Uricchio, T., Ballan, L., Bertini, M., Del Bimbo, A.: An evaluation of nearest-neighbor methods for tag refinement. In: Proc. of ICME (2013)
13. Villegas, M., Paredes, R., Thomee, B.: Overview of the imageclef 2013 scalable concept image annotation subtask. In: CLEF working notes, Valencia, Spain (2013)