Proc. ACM Int. Conf. on Multimedia Retrieval (ICMR 2015), Shanghai, China, June 2015.

Exploiting Multiple Web Resources towards Collecting Positive Training Samples for Visual Concept Learning

Olga Papadopoulou
Information Technologies Institute / CERTH
Thermi 57001, Greece
olgapapa@iti.gr

Vasileios Mezaris
Information Technologies Institute / CERTH
Thermi 57001, Greece
bmezaris@iti.gr

ABSTRACT

The number of images uploaded to the web is enormous and is rapidly increasing. The purpose of our work is to use these for acquiring positive training data for visual concept learning. Manually creating training data for visual concept classifiers is an expensive and time consuming task. We propose an approach which automatically collects positive training samples from the Web by constructing a multitude of text queries and retaining for each query only very few top-ranked images returned by each one of the different web image search engines (Google, Flickr and Bing). In this way, we sift the burden of false positive rejection to the Web search engines and directly assemble a rich set of high-quality positive training samples. Experiments on forty concepts, evaluated on the ImageNet dataset, show the merit of the proposed approach.

Keywords

Visual concept detection, Learning from Web data, Multiple text queries, Automatic training set construction

1. INTRODUCTION

Concept detection in images and video is a topic that has received significant attention. Most concept detection approaches proposed in the literature use manually-annotated data samples as training data for a machine learning method. The use of such manually-generated training data is also predominant in competitions such as TRECVID [1], which is organized annually for video annotation. However, manual image or video annotation with concept labels is a costly task, and its cost increases while the need of new concepts arises. This has inspired several recent works to investigate new techniques of collecting annotated image or video samples from the web, to use them as training data. Web images are often annotated with tags; however, these in many cases do not describe the visual content of the image [2, 3]. This happens due to different annotation criteria of the human Web users who annotate their images subjectively and with tags that are often meaningless when seen outside of personal context, or irrelevant to the target concepts of a visual classifier.

While most literature works try to overcome the above problem by downloading a large pool of Web images and subsequently applying to them some proprietary filtering or ranking algorithm, aiming to single-out those which truly contain the target concept, we are proposing a new approach which shifts the burden of the latter process to the Web. Our approach is motivated by the shape of the precision curve of Web image search results. Specifically, the precision of the search results for the first few top-ranked images (e.g., the first 50) is typically very high (above 90%), and decreases while the number of returned images (i.e. recall) increases. This means that by limiting ourselves to collecting just these few top results of a Web search engine, which were ranked as top results by considering a vast amount of Web resources that we could not possible take advantage of in a proprietary post-query selection process, we can render obsolete the latter process and still obtain high-quality training data. The above process could have the downside, though, that the size of the collected training corpus would be very small. To alleviate this, we propose a multi-query approach that automatically formulates multiple queries that are related to a single visual concept and returns the top-ranked images for each of them.

The rest of the paper is organized as follows. Section 2 reviews the related literature. In Section 3 the proposed framework is presented. A detailed description of the proposed approach is given in Section 4, and the experimental results are presented in Section 5. Finally, conclusions are drawn in Section 6.

2. RELATED WORK

The acquisition of positive and negative training examples is a vital step towards training a visual concept detector. Two main directions are followed for collecting the required training data: i) manual labeling of training data and ii) automatic collection of training samples from the Web. The first direction is extensively used in the literature and in benchmarking activities related to concept detection, such as TRECVID [1] and ImageNet [4]. However, the significant effort and scalability limitations arising from the need to manually annotate a high number of images or videos for every different concept that one needs to train a detector for, have motivated several recent works to explore the possibility of collecting training data automatically from the Web. The procedure that most of the related methods follow in

ICMR'15, June 23–26, 2015, Shanghai, China.

http://dx.doi.org/10.1145/2671188.2749338.

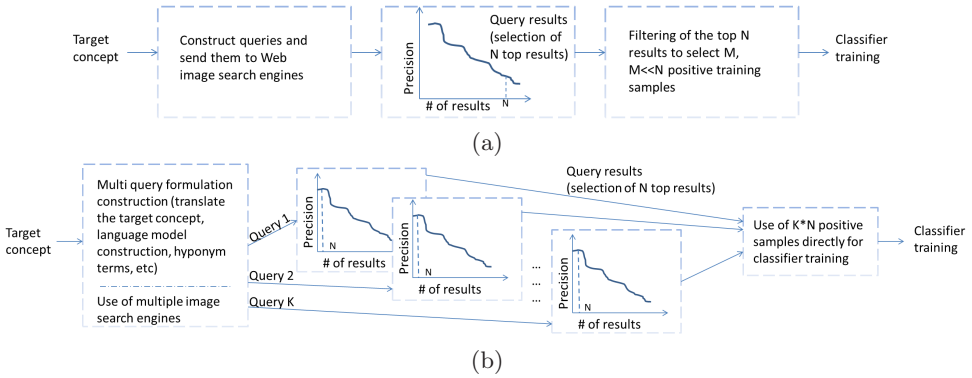

Figure 1: (a) Typical approach for collecting positive training samples from the Web, for visual concept detector training. (b) Proposed approach.

this direction is illustrated in Fig. 1(a): i) they construct one or a few queries that relate to the target concept, and they query the Web using one of the available image search engines, ii) they download a large number N of images that are returned for each query, iii) they apply a filtering or ranking algorithm on these collected images so as to discard the ones that, based on various criteria, are believed to be false positives. Specifically, in [5], given a concept, the WordNet ontology is used for producing a concept hierarchy, with the target concept as root node. Each node (concept) in the hierarchy is used for querying a search engine (Flickr) and download the top-ranked images. After that, the downloaded images of each node are pruned using a method called Semantic Field (SF). The final set of positive training samples for the target concept is constructed by pooling examples from all the nodes in the hierarchy. Similarly, in [6] a large amount of training examples are initially downloaded for each formulated query and are then pruned, with the difference that a text-based Web search engine is used for querying. Other works rely on a small, highly accurate set of positive training samples to start with and an incremental approach for obtaining the final training corpus, e.g., Li et al. [7].

Apart from collecting positive training examples from the Web, some works deal with the selection of negative examples. For this, random sampling is the most commonly used approach [8].

3. PROPOSED FRAMEWORK

An overview of the proposed approach is given in Fig. 1(b). In contrast to the typical approach of the literature (Fig. 1(a)), we shift the burden of selecting a high-quality set of positive samples to the Web image search engines, by exploiting the fact that when N receives a sufficiently low value, the N top results of a web query are almost always correct. Doing this when using a single query, however, significantly limits the number of positive samples that we would collect; to alleviate this, we adopt instead a multi-query approach which formulates K queries, i) by translating the target concept to different languages, ii) combining the target concept term with related WordNET terms, and iii) similarly combining the original target concept with terms extracted by Web text search. In this way we get only N images per query but overall several hundreds or even thousands of images per concept.

Performing some tests on the results of Web image search engines, we confirmed that for low values of N, the top$-N$ results are almost always correct. For instance, we queried the Google image search engine for concept *bird* and the Flickr image search engine for concept *animal*, and manually evaluated the top 1000 returned results. Going at a depth of $\simeq 600-900$ images (as in, for instance [5, 6]), the average precision is between $\simeq 55\% - 80\%$ in our example queries. On the contrary, at a depth of only e.g., 50 images, a very clean set of training samples (average precision $> \simeq 95\%$) can be directly retrieved.

The advantages of the proposed approach are reduced computational complexity, since no post-processing (i.e. pruning) is applied to the Web search results, and accuracy of resulting concept detectors, which is due to having a set of well-annotated positive training samples.

4. FORMULATING MULTIPLE CONCEPT-RELATED QUERIES

Given a target concept, three sets of queries are formulated in order to query Google, Bing and Flickr image search engines and select positive training examples. In Fig. 2, the proposed method's steps are shown, and a detailed description of each query set is presented below.

"Translation" set of queries

The target concept is first automatically translated into fifteen languages[1] and the translated terms are used to query the Google and Bing image search engines. In each case the Google or Bing domain of the specific country is called, in order to avoid receiving duplicate results; for example, for German: google.de and de.bing.com. Flickr image search does not provide such functionality, thus it is not used in this query set.

The maximum number of images that we can collect by applying the *Translation set* of queries for a target concept is calculated by $T_{tr} = E * N * L$, where E is the number of different image search engines that we use ($E = 2$), N is the number of top-ranked images that we retain and L is the number of different languages ($L = 15$). In practice, despite the translation and the use of different country

[1]German, French, Greek, Italian, Spanish, Chinese, Indonesian, Russian, Romanian, Bulgarian, Danish, Dutch, Finish, Hungarian, Portuguese; using Google translate

domains during Web search, we will still receive some duplicate images. In tests with a few example concepts and for $N = 24$ (720 images downloaded in total per concept) we saw that the amount of duplicate images per target concept varied between 5% and 45% of T_{tr}.

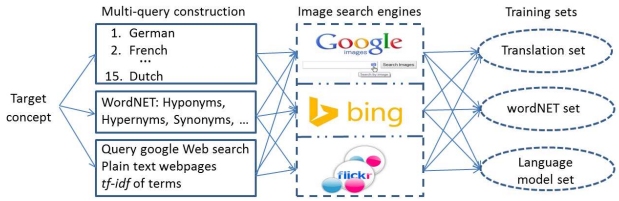

Figure 2: Training data collection for a target concept

"wordNET" set of queries

The second set of queries is built using the WordNET [9] lexical database. WordNET groups English words into sets, such as synonyms, hyponyms, hypernyms, which are called *synsets*. For each target concept, we get from WordNET a list of such related terms. Our queries to the Web image search engines are then formed by combining each of the returned terms with the target concept using AND (&) (i.e. target_concept & wordNET_term).

Although the extracted terms should be semantically related to the target concept, this is not always the case. In such a case the images that will be returned by this term combination will not visually depict the target concept. For example, a *synset* term returned for target concept "dog" is "hot dog". Images returned by the "dog & hot dog" term combination will often not depict the animal dog. Such noisy results should not be included in our training corpus, thus should not be retrieved at all. In order to achieve this, we calculate the similarity of each wordNET-extracted term with the target concept by applying: i) the *umbc* similarity measure [10], which combines the use of thesaurus (e.g., WordNET) and statistics from a large corpus for computing word similarity and ii) the *easyesa* similarity measure, [11], which uses the Wikipedia commonsense knowledge base for a statistical analysis of the co-occurrence of words in the text. Based on experiments on five concepts and their related terms extracted by WordNET, we heuristically chose the value of thresholds $t_u = 0.5$ and $t_e = 0.05$, for the above similarity measures, and we perform the term combination queries described above only when the *umbc* and *easyesa* similarity scores of the two terms being combined are above the corresponding threshold. We assume that these threshold values are suitable for all concepts and we use them throughout our experiments with forty concepts in Section 5.

"Language model" set of queries

The third set of queries is based on text Web search results. The web is flooded with text, and we use it to retrieve more terms related to the target concept and build some sort of a language model for the latter. These terms are combined with the target concept, similarly to what was done in the previous subsection, to construct more text-queries. The steps that we follow are: Given a target concept i) a query of the target concept term is sent to the Google text search engine, ii) the W first Web pages are retrieved, iii) the text of the retrieved Web pages is extracted, iv) term frequency inverse - document frequency (td-idf) [12], which reflects how important a word is to a document in a collection or corpus, is applied on the extracted plain text, resulting in a vector of related or not-related terms. Terms with td-idf higher than t_τ are considered related to the target concept. We set $t_\tau = 0.01$ by conducting similar experiments to those described above for t_u and t_e.

However, there are cases where a term appears frequently in the text of a Web page returned by querying the target concept but is not visually related with it. For example, the returned terms with highest tf-idf for target concept *airplane* are: airplane, plane, aircraft, jet, paper airplane, rc airplane, flight, cockpit, flying, **passenger**, **pilot**, **passenger cabin**,... Several terms, such as the ones in bold, are related to the target concept but do not visually describe it. To reject these terms, we again further assess the suitability of each term combination using measures *umbc* and *easyesa*, together with their previously set thresholds (t_u, t_e). The retained terms are combined (AND (&)) with the target concept to generate additional queries.

5. EXPERIMENTAL RESULTS

We experimented with a pool of forty concepts and an evaluation set of 17881 images from ImageNet dataset [4]. The positive training samples for training our concept detectors are collected from the Web using three image search engines and the multi-query formulation approach, as described in Section 4. Negative sample selection is out of the scope of our study; thus, the commonly-used solution of random selection is adopted and approximately 5000 images are used as the negative training samples. The same negative samples are used throughout the experiments.

For training concept detectors, we used the SIFT, RGB-SIFT and opponentSIFT local descriptors and the methodology of [13]. The trained concept detectors output, when assessing a new image, is a score in the range [0,1], where higher values indicate higher confidence that the image in question depicts the concept.

For evaluating a trained concept detector, this detector is applied to the entire evaluation set, the set's images are ordered according to the detector's output score (in descending order) and Average Precision (AP) [14] is calculated; Mean Average Precision (MAP) is also calculated as the mean AP value across all concepts.

Use of the different training sets

The proposed multi-query approach collects positive training samples using three sets of queries, as detailed in Section 4. We evaluate the use of the collected positive examples in two ways:

Early fusion (union of the training sets): We consider the images returned by all constructed queries for a given concept as a single training set, on which the detector is trained.

Late fusion (individual training sets): We consider the images returned by each of the three query sets as a separate training set; for the given concept, a different detector is trained on each and the three detectors are combined by late fusion (averaging of output scores).

In Table 1, the concept detection results are presented for the two above cases. These results show that *Late fusion* consistently produces better results than *Early fusion*, and also better than using alone any one of the training sets

assembled using a single set of queries. Therefore, the *Late fusion* approach is adopted in the rest of the experiments.

Optimal number of top-ranked images to download

The above experiment was run for different values of N, in order to also examine the optimal number of top-ranked images to download per query and target concept. From the results of Table 1 we can see that the optimal is $N = 24$. The MAP increases while the value of N increases from 8 to 24, because more positive examples are included in the training corpus without introducing significant noise, but then starts to decrease, as a result of the set of positive training samples becoming more noisy.

Table 1: MAP for different training sets and the number of retained top-ranked images $N = 8, 16, 24$ and 32

Training set\ N value	Top8	Top16	Top24	Top32
Translation set	0.347	0.363	0.37	0.316
wordNET set	0.308	0.331	0.35	0.34
Language model set	0.281	0.305	0.33	0.32
Early fusion	0.35	0.36	0.37	0.35
Late fusion	0.387	0.399	**0.41**	0.4

Comparisons

We compared the proposed approach with the following:

Baseline: a query of the target concept term is sent to the Flickr image search engine and a fixed number X of examples is returned. We empirically set $X = 600$, as is often the case in the literature. All images returned are considered positive training examples, and no further pruning is applied on the images.

ImageNet examples: We downloaded the manually annotated images of the target concepts and trained the classifiers.

Zhu et al. [5]: We implemented the sampling approach introduced in [5] and trained SVM-based classifiers (following the methodology of [13]) with the selected examples.

Table 2: Comparison with baselines and SoA

Approach	MAP	Approach	MAP
Baseline	0.252	ImageNet examples	0.46
Zhu et al. [5]	0.3	Proposed approach	0.41

Table 3: Relevance fraction for a few concepts

Approach\Concepts	animal	bear	beach	boat	book
Baseline	0.79	0.57	0.69	0.62	0.47
Zhu et al. [5]	0.79	0.75	0.81	0.64	0.53
Translation set	0.91	0.87	0.96	0.77	0.87
wordNET set	0.82	0.6	0.96	0.91	0.85
Lang. model set	0.7	0.55	0.75	0.72	0.86
Proposed approach	0.86	0.8	0.84	0.82	0.87

In Table 2, the results of all conducted experiments, in terms of MAP, are presented. As expected, the Baseline approach is the one that performed the worst, with MAP equal to 0.252, since its training set contained many noisy samples. The ImageNet manually-generated training set resulted in the best trained concept detectors, with MAP equal to 0.46. The implemented state-of-the-art approach, Zhu et al. [5], performed better than the Baseline, achieving a MAP of 0.3. Our proposed approach with a MAP of 0.41 significantly outperforms the baseline and [5].

To explicitly assess whether false positives are included in the positive training samples that we collect with the proposed approach, we manually checked this for a small number of concepts. We quantify the results of this manual assessment by calculating the *fraction of relevance*, which is defined as the fraction of true positives in the collected image set. As shown in Table 3 the proposed approach collects fewer false positives than the compared approaches.

6. CONCLUSIONS

Web images are suitable candidates to serve as training examples for visual concept learning. We have presented an approach of multi-query formulation for collecting positive training examples from the Web. Top-ranked images that are returned by image search engines, as a result of automatically formulating and submitting a large number of queries, allow us to directly collect high-quality sets of positive training samples.

7. ACKNOWLEDGMENTS

This work was supported by the EC under contracts FP7-600826 ForgetIT and FP7-287911 LinkedTV.

8. REFERENCES

[1] P. Over, G. Awad, et al., "Trecvid 2014 – an overview of the goals, tasks, data, evaluation mechanisms and metrics," in *Proc. TRECVID 2014*. NIST, USA, 2014.

[2] X. Li, C. G. Snoek, et al., "Harvesting social images for bi-concept search," *IEEE Trans. on Multimedia*, vol. 14, no. 4, pp. 1091–1104, 2012.

[3] X. Li, C. G. Snoek, and M. Worring, "Unsupervised multi-feature tag relevance learning for social image retrieval," in *Proc. Int. Conf. on Image and Video Retrieval*. ACM, 2010, pp. 10–17.

[4] J. Deng, W. Dong, et al., "Imagenet: A large-scale hierarchical image database," in *Proc. Int. Conf. on Computer Vision and Pattern Recognition*. IEEE, 2009, pp. 248–255.

[5] S. Zhu, C.-W. Ngo, and Y.-G. Jiang, "Sampling and ontologically pooling web images for visual concept learning," *IEEE Trans. on Multimedia*, vol. 14, no. 4, pp. 1068–1078, 2012.

[6] F. Schroff, A. Criminisi, and A. Zisserman, "Harvesting image databases from the web," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 33, no. 4, pp. 754–766, 2011.

[7] L.-J. Li and L. Fei-Fei, "Optimol: automatic online picture collection via incremental model learning," *Int. Journal of Computer Vision*, vol. 88, no. 2, pp. 147–168, 2010.

[8] X. Li and C. G. Snoek, "Visual categorization with negative examples for free," in *Proc. 17th ACM Int. Conf. on Multimedia*, 2009, pp. 661–664.

[9] C. Fellbaum, *WordNet: An Electronic Lexical Database*, Bradford Books, 1998.

[10] L. Han, A. Kashyap, et al., "Umbc ebiquity-core: Semantic textual similarity systems," in *Proc. 2nd Joint Conf. on Lexical and Computational Semantics*, 2013, vol. 1, pp. 44–52.

[11] D. Carvalho, C. Callı, et al., "Easyesa: A low-effort infrastructure for explicit semantic analysis," in *Proc. 13th Int. Semantic Web Conference (ISWC)*, 2014.

[12] J. Ramos, "Using tf-idf to determine word relevance in document queries," in *Proc. 1st Instructional Conf. on Machine Learning*, 2003.

[13] F. Markatopoulou, N. Pittaras, et al., "A study on the use of a binary local descriptor and color extensions of local descriptors for video concept detection," in *Proc. MultiMedia Modeling, Springer*, 2015, pp. 282–293.

[14] A. F. Smeaton, P. Over, and W. Kraaij, "Evaluation campaigns and trecvid," in *Proc. 8th ACM Int. Workshop on Multimedia Information Retrieval*, 2006, pp. 321–330.