Multimed Tools Appl
DOI 10.1007/s11042-016-4142-3

Query expansion for object retrieval with active learning using BoW and CNN feature

Xin Zhao[1] · **Guiguang Ding**[1]

Received: 16 April 2016 / Revised: 13 September 2016 / Accepted: 8 November 2016

Abstract Most effective particular object and image retrieval approaches are based on the bag-of-words (BoW) model, and all state-of-the-art performance mainly involves a query expansion procedure, which is able to significantly improve retrieval results. Nowadays, Convolutional Neural Network(CNN) is widely applied in computer vision field, including image classification, caption, recognition and retrieval, etc. We introduce an extension to query expansion: an automatic method to select good candidate samples for interactive annotation which is used in query expansion using both BoW method and CNN feature. In this work, we address the query expansion framework using active learning, where the main focus is on the sample selection step in the process of query expansion. More specifically, we propose an active sample selection algorithm based on binary relevance classification, based on the assumption that most confusing samples of the classifiers have high probability to contain helpful true positives for query expansion, which significantly improves the retrieval performance. It takes full use of the multimodal information of the shortlist obtained from the basic retrieval to train a binary relevance classifier, which is used to pick up the most confusing samples for human annotation, with top list as unlabeled data and bottom list as fake negatives. And it can achieve a faster and better retrieval than naive top sample selection method. We also fuse BoW vector and CNN prediction in the retrieval system for a better performance. To evaluate the performance of our proposed method, experiments are conducted on Standard Oxford (5K and 105K) and Paris (6K) datasets, and experimental results and comparison with the state-of-the-art methods demonstrate the effectiveness of the proposed method.

✉ Guiguang Ding
dinggg@tsinghua.edu.cn

Xin Zhao
zhaoxin19@gmail.com

[1] Beijing Haidian Disctrict, Tsinghua University, Beijing Shi, China

Published online: 28 December 2016

Keywords Active learning · Query expansion · Convolutional neural network · Image retrieval

1 Introduction

Many successful particular object and image retrieval approaches are based on the bag-of-words (BoW) model proposed in [24]. In such retrieval methods, the BoW description is first generated for both the query and the images in the dataset, respectively. Then, the search engine can measure the distance between the query and each image, and a shortlist of potentially relevant images is efficiently retrieved by weighted BoW vectors using inverted index or any other sparse vector algorithm. The BoW type retrieval has been widely investigated in recent years, such as feature detectors and descriptors [16–18, 27] , vocabulary construction [9, 19, 22, 24], spatial verification and re-ranking [9, 22], document metric learning [3, 10, 12] and dimensionality reduction [11, 21].

There are a lot of studies that apply Convolutional Neural Network in image retrieval domain. An online multimodal deep similarity learning method is proposed in [26, 28]. They put similarity learning network after traditional CNN classification network and use triplet hinge loss to refine the whole network, which significantly improves the retrieval performance compared to using CNN fully-connected layer feature directly. It is noted that CNN feature without refining performs not so well as BoW feature in the image retrieval task. And the refining method in [26, 28] requires relevant triplet training data which is hard to obtain. So we propose a method using CNN prediction to boost BoW retrieval.

It is noted that query expansion(QE) [6] brought a significant boost for retrieval performance [6, 10, 20, 23], and nearly all state-of-the-art retrieval results involve a query expansion step. Generally, QE is composed of two procedures, i.e., sample selection and expanded query modeling. Existing QE methods mainly enrich the query model by adding features of spatially verified images automatically. Most of QE methods focus on expanded query modeling step, where the most common modeling method is weighted averaging every expanded samples such as average query expansion(AQE) [6] and discriminative query expansion(DQE) [1]. It is noted that sample selection plays an even more important role in QE, where a good sample selection method can introduce more informative data and less noise. Existing QE methods use strong spatial verification on the candidate images in the shortlist to assure the selected images are relevant to the query in the sample selection step. It has been observed in [6] that if the shortlist has enough true positives, the spatial verification can always correctly identify relevant images, and, the expanded query is significantly better than the original single query. On the contrary, if there are no, or very few, relevant images in the shortlist, the query expansion method with spatial verification cannot work well.

There are two main problems in existing spatial verification query expansion method: (i) the selection of the threshold of spatial verification, which is hard to find and may vary for different topics. (ii) spatial verification method tends to pick up samples that are similar with the query in BoW space, with which we can hardly introduce true positives from different views of the query, which usually contains more information.

To tackle the problems of sample selection with spatial verification, we introduce an interactive sample selection framework which can explore better samples for QE.

Two naive sample selection methods are Random Sample Selection(RSS) and Top Sample Selection(TSS). TSS selects top-k samples from the shortlist. Top samples from the shortlist are similar to the original query and more likely to be positive samples. So TSS

sometimes picks up too many positive samples for QE which increase less information but introduce more noise from background. As it is known, the expenditure of time of BoW method is linear growth with the number of query words. So TSS always costs more query time especially in big dataset. RSS randomly selects k samples from the shortlist. This method performs well with a big density of true positives in the shortlist. Otherwise, if the shortlist is sparse, it can hardly pick up new positives. We take TSS and RSS as baseline sample selection methods for query expansion.

We first train an binary relevance classifier with the basic retrieval shortlist which we will state in Section 3 and then pick up the most confusing samples using both BoW ranking and the CNN feature of selected images for human annotation. Finally, we expend the selected positives annotated by human into original query. This method is called Active Sample Selection (ASS). It can expand less but enough useful samples in our query model and lead to a faster and better QE retrieval. The proposed method has been evaluated on Oxford5K, Oxford105K and Paris6K datasets. Experimental results and comparison with state-of-the-art methods demonstrate the effectiveness of the our proposed method.

The rest of the paper is structured as follows. Related work is provided in detail in Section 2. The proposed retrieval method using both BoW vector and CNN prediction is presented in Section 3. The proposed Active Sample Selection method is presented in Section 4. Section 5 provides experimental results and comparison with existing methods. We finally conclude this paper in Section 6.

2 Related work

2.1 Query expansion

Query expansion is proved to be a powerful strategy to improve the performance in both text retrieval and content based image retrieval (CBIR). The search engine takes use of the result shortlist of the original query to issue a new ”expanded” query. In this way, additional relevant terms can be added to the query.

The query expansion was first introduced into the visual domain by [6]. A strong spatial constraint between the query image and dataset makes verification accurate, wiping out false positives that typically ruin text-based query expansion as much as possible. These verified images can be used to make up a new expanded query.

In [6], a number of query expansion strategies were proposed. All of them follow a similar pattern: images in a shortlist are spatially verified against query features, images with sufficient numbers of matches (inliers) are back-projected by the estimated affine transformation into the query region, and, finally, a new query is issued. The proposed strategies in [6] concentrate on multi-round query expansion or ”clean” features (interest points) selection.

The simplest well performing query expansion method is called average query expansion(AQE). The expanded documents are simply averaged to get a new query. This method is both quick and popular. We take AQE as our fusion method of hand-annotated true positives.

2.2 Spatial verification

LO-Ransac [4, 14] is a fast spatial verification algorithm which is widely used in state-of-the-art image retrieval system. It can estimate an affine transformation from query to dataset

image and return the number of interest point inliers. [6] uses LO-Ranscac to spatial verify top images from the shot list. In [5], an incremental spatial re-ranking (iSP) method was proposed to improve the performance of spatial verification. However, the spatial verification algorithm is much more slower than the BoW scoring. It is common to apply spatial verification only on a few top images.

2.3 Active Learning

Active learning methods aim to find samples with more information(or to be easily confused) and present them to human for annotation, so that active learning system can improve its performance from less interactive annotation.

The simplest and most commonly used active learning framework is uncertainty sampling [15]. In this frame work, an active learner queries the instances which are least certain to label.

In [25], SVM active learning method is used in image retrieval domain. They achieve a retrieval system as below. Firstly, system randomly select k samples for human annotation. And then an SVM classifier is trained by the labeled samples. Next, system select top-k relevant samples for human annotation, and train new classifiers round by round. Their method needs a multi-round relevance feedback process to achieve a good retrieval result, and therefore need more human annotation.

ASS method in this work is to pick up "good " samples for query expansion, which are not so similar to the original query without multi-round relevance feedback. They should be true positives with a high possibility.

3 Boost BoW retrieval with CNN prediction in landmark retrieval

In the basic BoW retrieval, we found that many confusing negative samples brought to the top of the result list are even not buildings. They are taken because they contain lots of sift points in different clusters. Mostly, the confusing negative samples are even not buildings and they are "similar" to many of the queries. If we can wipe out those samples, we can get a better retrieval result.

CNN has solved image classification problem successfully. It can give a correct image classification easily without too much training. So we can use CNN classifier to wipe out the confusing negative samples mistaken by BoW retrieval. Text based image search engine performs so well nowadays that we can obtain text labeled image easily. We can use web data as train set to fine-tune AlexNet to be a building binary classifier. We checked the web data used for fine-tuning to make sure it does not contain any images in target dataset. The output layer after soft of our network is a 2-dimension float vector v_j, which is a probability output to judge if X_j is a building ($P_{building}(X_j) = v_j[building]$).

The distance of query Q_i and dataset pattern X_j is

$$dis(Q_i, X_j) = dis_{AS}(Q_i, X_j) - \gamma P_{building}(X_j),$$

where $dis_{AS}(Q_i, X_j)$ is the asymmetrical dissimilarity [29] between Q_i and X_j.We use ℓ_1 metric for asymmetrical dissimilarity. This method is referred as CNN Prediction Boosting(CNNPB).

4 Query expansion with active learning

4.1 Query expansion with active learning framework

We use active learning in the sample selection step of query expansion. The whole system is divided into two parts: (i) BoW object retrieval system, (ii) active query expansion system. The query is the input of BoW object retrieval system, and we can obtain an original shortlist from it. Both the shortlist and the query are fed to active query expansion system. The active query expansion system first trains a binary relevant classifier with the CNN feature of the input shortlist to pick up the most uncertain samples, which are not so similar to the query for human annotation. The selected samples should be positives with a high possibility. Sample Selection step takes use of BoW ranking and the CNN feature of top result. Then the selected true positives are used to model a new expanded query, which is a feedback to the BoW object retrieval system for a second round retrieval. Finally, the result is presented to user. The query expansion with active learning framework is as Fig. 1.

4.2 Positive and unlabeled learning(PU-learning)

Our idea was inspired by the former works about binary classification from only positive and unlabeled data(PU classification). A naive approach to PU classification is to train a classifier to separate positive and unlabeled samples. However, such a naive approach yields a poor solution since the unlabeled dataset consists of both positive and negative data. And the PU classifier trained in this way has a systematic estimation bias. In [2], the contamination model of the marginal distribution of an unlabeled pattern X was proposed:

$$X \sim p_X = (1-\pi)p_{-1} + \pi p_1,$$

which leads following expected misclassification rate E_X

$$E_X = (1-\pi)E_{-1} + \pi E_1, \tag{1}$$

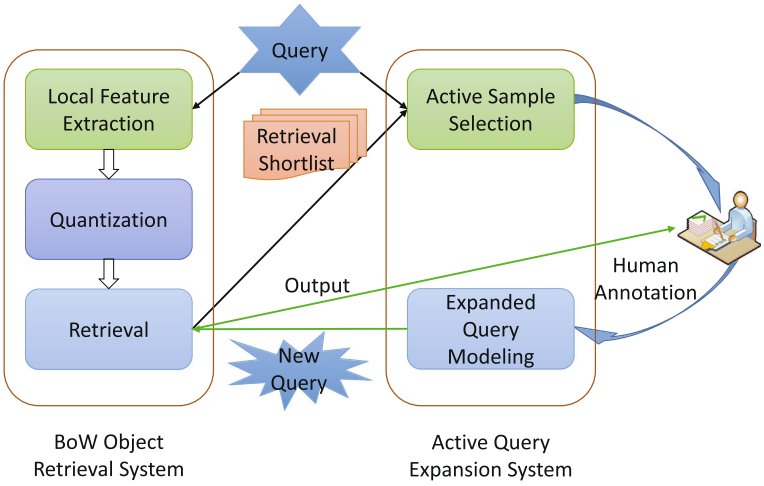


Fig. 1 Query expansion with active learning framework

where π is the density of positive samples in the unlabeled set. Based on the work of [2], a number of non-convex and convex loss functions were proposed in [7, 8]. Above these loss functions, [8] noticed that the objective function with squared loss of classification has an analytical solution, which provides a fast training process we need in retrieval task.

We can apply PU classification method in the active learning process to solve the problem of lack of training data. We have got a query which is definitely a true positive. And, we can obtain an unlabeled set, a mixture of positives and negatives from the top of retrieval shortlist, which increases our information about positive samples. Our learning problem is with very small positive set and large unlabeled set, which can be solved by PU-Learning.

In our ASS method, we adopt (1) and square loss function.

4.3 Active sample selection

Our method of Active Sample Selection is to train a classifier, which is to judge if a candidate sample is relevant or not. This classifier will select the most uncertain samples for human annotation, because the more confusing a sample is, the more information will be added by the annotation of the sample. PU classification is helpful to the classification problem we are facing. We train the classifier with CNN feature because it has much lower dimension than Bow feature, so that the time cost is smaller.

Formulation of PU classification Let $x \in R^d$ be a d-dimensional pattern and $y \in \{1, -1\}$ be a class label. We assume that we have a positive dataset χ and an unlabeled dataset χ' i.i.d. as

$$\chi := \{x_i\}_{i=1}^{n} \sim p(x|y=1),\ \chi' := \{x_i'\}_{i=1}^{n'} \sim p(x),$$

where $p(x|y)$ is the class-conditional probability density of patterns and $p(x)$ is the marginal probability density of patterns. Since the unlabeled dataset χ' consists of positive and negative samples, the marginal density is $p(x) := \pi^* p(x|y=1) + (1-\pi^*)p(x|y=-1)$. The goal is to learn a classifier $g(x)$ that assigns a label $\hat{y} = sign(g(x))$.

The optimal classifier g^* is given by $g^* = argmin_g \mathcal{J}(g)$, where $\mathcal{J}(g)$ is the expected misclassification rate when the classifier $g(x)$ is applied to unlabeled samples distributed according to $p(x)$:

$$\mathcal{J}(g) = \pi^* E_1[l(g(X))] + (1-\pi^*)E_{-1}[l(-g(X))], \tag{2}$$

where $l(\cdot)$ is the loss function, π^* is the density of positive samples in the all sample space.

PU classification for loss minimization In [2], the contamination model of the marginal distribution of an unlabeled pattern X was proposed. And the expectation loss of X is as (1). E_{-1} cannot be estimated directly, so according to (1) we can translate (2) to:

$$\mathcal{J}(g) = \pi^* E_1[\hat{l}(g)] + \frac{\pi^*}{\pi} E_X[l(-g)] + (1 - \frac{\pi^*}{\pi})E_{-1}[l(-g)], \tag{3}$$

where $\hat{l}(g) = l(g) - l(-g)$.

Under the assumption that the density of positive samples in the all sample space and the unlabeled set is the same, π is equal to π^*. (3) can be simplified to:

$$\mathcal{J}(g) = \pi^* E_1[\hat{l}(g(X))] + E_X[l(-g(X))], \tag{4}$$

which is a PU classification objective function.

Using fake negative samples Without the assumption of $\pi = \pi^*$, we must use (3) as objective function. Fortunately, it is observed that the bottom images from the retrieval shortlist are negative samples with a great probability. We can use the bottom images as fake negative samples in (3). As the formulation of PU classification, the fake negative set is represented as χ'' i.i.d as

$$\chi'' := \{x_i''\}_{i=1}^{n''} \sim p(x|y=-1).$$

Figure 2 shows several top images and bottom images of some queries in the shortlist.

Squared loss The squared loss, defined as $l_S = \frac{1}{4}(z-1)^2$, results in the following objective function:

$$\mathcal{J}_S(g) = -\pi^* E_1(g) + \frac{\pi^*}{\pi} E_X[(g+1)^2] + (1 - \frac{\pi^*}{\pi}) E_{-1}[(g+1)^2]. \tag{5}$$

We choose a linear $g(x)$, $g(x) = \alpha^\top x + b$. (5) can be estimated as

$$\begin{aligned}\mathcal{J}_S(\alpha, b) = &-\frac{\pi^*}{n}\sum_{i=1}^{n}\alpha^\top x_i - \pi^* b + \frac{\pi^*}{4n'\pi}\sum_{i=1}^{n'}(\alpha^\top x_i' + b + 1)^2 \\ &+(\frac{1}{4n''} - \frac{\pi^*}{4n''\pi})\sum_{i=1}^{n''}(\alpha^\top x_i'' + b + 1)^2 + \frac{\lambda}{2}(\alpha^\top\alpha + b^2).\end{aligned} \tag{6}$$

To solve (6), let

$$\varphi(x) = \begin{bmatrix} x \\ 1 \end{bmatrix}, \alpha'^\top = \begin{bmatrix} \alpha^\top \\ b \end{bmatrix},$$

$$\Phi(X^*) = [\varphi(x_1^*), \varphi(x_1^*), \ldots \varphi(x_{n^*}^*)],$$

where X* represents positive pattern, unlabeled pattern or negative pattern.

We can get

$$\begin{aligned}\mathcal{J}_S(\alpha') = &-\frac{\pi^*}{n}\sum_{i=1}^{n}\alpha'^\top\varphi(x_i) + \frac{\pi^*}{4n'\pi}\sum_{i=1}^{n'}(\alpha'^\top\varphi(x_i') + 1)^2 \\ &+(\frac{1}{4n''} - \frac{\pi^*}{4n''\pi})\sum_{i=1}^{n''}(\alpha'^\top\varphi(x_i'') + 1)^2 + \frac{\lambda}{2}\alpha'^\top\alpha',\end{aligned} \tag{7}$$

Fig. 2 Some top and bottom images of these two queries. Green for positive and red for negative

which can be written into

$$\hat{\mathcal{J}}_S(\alpha') = \frac{1}{2}\alpha'^{\top}[\frac{\pi^*}{2n'\pi}\Phi_U^{\top}\Phi_U + \frac{\pi-\pi^*}{2n''\pi}\Phi_N^{\top}\Phi_N + \lambda I]\alpha'$$
$$+(\frac{\pi^*}{2n'\pi}1^{\top}\Phi_U + \frac{\pi-\pi^*}{2n''\pi}1^{\top}\Phi_N\alpha' - \frac{\pi^*}{n}1^{\top}\Phi_U)\alpha' + C.$$

Algorithm 1 Active sample selection (Part I)

```
procedure CLASSIFIER TRAINING(Q, L)
    Input: query Q, shortlist of BoW L
    X_P = Q;
    X_U = L[top(n')];
    X_N = L[bottom(n'')];
    Φ_P := [X_P, 1]; Φ_U := [X_U, 1]; Φ_N := [X_N, 1];
    Caculate α'as(8);
    return α', Φ_U;
end procedure
```

Algorithm 2 Active sample selection (Part II)

```
procedure SAMPLE SELECTION(α', Φ_U)
    Input: classifier α', unlabeled set Φ_U
    toplist = getTopList(Φ_U, k_1);
    Score = abs(α'^⊤ Φ_U);
    selectlist = getMinScoreList(Φ_U, Socre, k_2);
    result = Union(toplist, selectlist);
    return result;
end procedure
```

The minimizer of this objective function can be analytically obtained as

$$\alpha' = M^{-1}[\frac{\pi}{n}\Phi_P^{\top}1 - \frac{1}{2n'}\Phi_U^{\top}1 - \frac{\pi-\pi^*}{2n''\pi^*}\Phi_N^{\top}1], \tag{8}$$

where

$$M = \frac{1}{2n'}\Phi_U^{\top}\Phi_U + \frac{1}{2n''}(\frac{\pi}{\pi^*} - 1)\Phi_N^{\top}\Phi_N + \lambda' I, \lambda' = \frac{\pi}{\pi^*}\lambda.$$

The square loss leads to an objective function with analytical solution, which makes training process fast. And this method can speed up by feature dimension reduction. The PCA-based feature dimension reduction method can be an easy extension to it.

4.4 Active sample selection algorithm

Our Active Sample Selection method includes two steps: (i) classifier training, (ii) sample selection. In the classifier training step, we pick up n' samples from the top of the BoW shortlist as unlabeled set and n'' samples from the bottom as fake negative set. The procedure outputs a linear classifier α' as Algorithm 1. shows.

In the sample selection step, we classify the unlabeled set with α' and union the original top k_1 list and the k_2 min score list from the unlabeled set as the final result. The sample selection algorithm is as Algorithm 2.

Then we model the expanded query with AQE, and feed it back to the BoW retrieval system and get the final retrieval result.

5 Experiment evaluation

5.1 Dataset and features

We evaluate our experiments on Oxford Buildings [22] and Paris [23] datasets. Additional 100k confuser images of the 75 most popular Flickr tags are provided with the Oxford Buildings [22]. For each of the Oxford and Paris datasets, the evaluation protocol defines 55 queries, five for each landmark, with precise ground truth. Each query has a bounding box of query object. The performance of all retrieval experiments is measured by the mean average precision(mAP).

The local feature used in BoW retrieval system is rootsift [1], which is quantized into an 1M AKM dictionary with idf weights. The dictionary from Oxford5K is used to quantize Oxford5K and Oxford105K. The 4096-dimensional CNN fc7 feature(the 7th fully-connected layer feature from CNN network) is extracted by a fine-tuned AlexNet [13], which is used for training the ASS classifier.

5.2 BoW retrieval system boosted by CNN prediction

Experiment Details Our baseline object retrieval system is based on query-adaptive asymmetrical dissimilarities(AS) [29]. It values more about query words with query-adaptive weights (Fig. 3).

Although all the queries are with precise bounding box, some of them are polluted by a few background words (sky, branches, grass, etc). Figure 4 shows some polluted queries. We wipe out all background words from the 55 queries. That is to say, $Q = I/O$, where Q is the final query, I is the word set in the bounding box, O is the word set outside.

The fusion parameter γ stated in Section 3 differs in different dataset. We do experiment on Oxford5K and Paris6K to pick up a proper γ value. Notice that the query in Oxford5K and Oxford105K is totally the same. So we take the same γ for them.

Experiment Result Figure 3. shows how γ influences the retrieval performance. The curve is with single-peak. The proper values are $\gamma = 10^4$ for Paris and $\gamma = 10^5$ for Oxford. The performance of our BoW retrieval system is listed in Table 1. CNN prediction can get

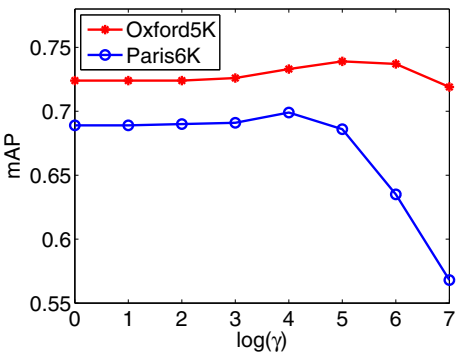


Fig. 3 The influence on retrieval performance(mAP) with different λ on Oxford5K and Paris6K

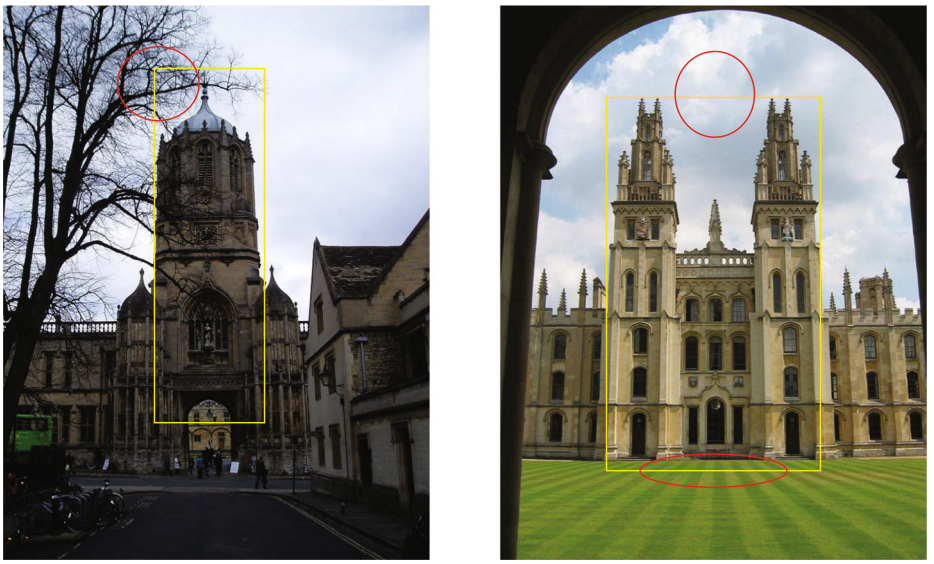


Fig. 4 Query in the bounding box polluted by the background words. The yellow rectangle is the bounding box of the queries. The red ellipse is the confuser words from background

a promotion around 0.01 in mAP. Our BoW retrieval system performs better than the basic tf-idf ranking system.

5.3 Sample selection for query expansion

Query Expansion The original query is cleaner than the dataset image, because we take advantage of the bounding box to wipe out some background words. So we should weight more for the original query when modeling the expanded query with AQE.

Sample Selection We pick up top 200 images and bottom 100 images from the original retrieval shortlist for ASS classifier training in our experiment. The parameter π and π^* denote the density of positive patterns in the unlabeled set and the full sample space. Although the density of different topics is different, we cannot get query-adaptive learning

Table 1 Some baseline performance without query expansion

Dataset	Oxford5k	Oxford105k	Paris6k
tf-idf weighting [22]	0.618	0.490	–
tf-idf weighting [5]	0.616	0.553	0.617
Asymmetrical dissimilarity(AS) [29]	–	0.739	–
AS without CNNPB	0.724	0.689	0.684
AS with CNNPB	0.739	0.697	0.699

The performance is evaluated by mAP. The weight learning parameter $\alpha_1 = 0.5$ in [29], and $\alpha_1 = 0.8$ in this paper

Fig. 5 Samples selected by ASS. Yellow samples are positives selected by top selection. Green samples are positives selected by the classifier. Red samples are negatives selected by the classifier. Blue circles represent occluded query object in negatives

weights. So we choose π and π^* close to mAP of the original retrieval list as an approximate value, which is $\pi = 0.7, \pi^* = 0.72$ in Oxford5K, Oxford105K and Paris6K.

Our experiment is about ASS, TSS and RSS on different annotation scale S. S refers to the number of samples presented to human for annotation. The annotation process by human can be simulated by looking up the groundtruth files. RSS randomly selects samples from the top 200 list for annotation. In the experiment of ASS, we pick up k_1 top samples to assure we can obtain some positive samples, while we select k_2 samples selected by the classifier for better query. We first study the influence of the ratio $r = k_1 : k_2$ for the performance of ASS. And then fix r to study the performance of ASS in different annotation scale S.

Figure 5. shows the samples selected by ASS for the some queries on Oxford5K, while $S = 10$. The samples selected by top selection(yellow) are more similar to the query. Compared with the query, the positives selected by the classifier(green) tend to be from different views, under different light condition or in different seasons, etc. And the negatives selected by the classifier(red) is confusing. Some of them(blue circle) even contain occluded query object.

Experiment Result Table 2. shows the different performance of ASS with different r when $S = 30$. The performance of ASS is sensitive to the value of r. When $r = 0$, the samples are all selected by the classifier. With the growth of r, the number of samples selected by TSS grows. The best r value varies for different datasets, because the information gain from the top list is different between different datasets. Samples at the top list are almost all positives but they tend to be similar. Samples selected by the classifier contain more

Table 2 mAP with different r when $S = 30$

Dataset	$r = k_1 : k_2$					
	0	0.2	0.5	1	2	5
Oxford5k	0.842	0.834	0.848	0.852	0.853	0.847
Oxford105k	0.795	0.784	0.801	0.811	0.813	0.807
Paris6k	0.743	0.734	0.733	0.727	0.726	0.719

ASS performs best when $r = 2$ on Oxford datasets and when $r = 0$ on Paris dataset

Table 3 Performance of query expansion methods with spatial verification or sample selection

Dataset	Basic	SP	TSS	RSS	ASS
Oxford5k	0.739	0.803	0.843	0.830	0.853
Oxford105k	0.697	0.758	0.804	0.794	0.813
Paris6k	0.699	0.713	0.716	0.732	0.743

Spatial verification is for top-30 images. $S = 30$ in sample selection methods. We pick up the best performance of different r for ASS

information but they may be sparse. In Oxford datasets, samples at the top list are not so similar with each other, so TSS brings improvement. In Paris dataset, top list samples are not so useful, so the best r in Paris is 0. The top list of Oxford datasets provides much more information of the query. We could see that TSS improves 0.102 in mAP from basic retrieval on Oxford5K while only 0.017 on Paris6K in Table 3.

We compare the experiment result of ASS with AQE with spatial verification, TSS and RSS when $S = 30$, which is listed in Table 3. ASS performs better than others at the same

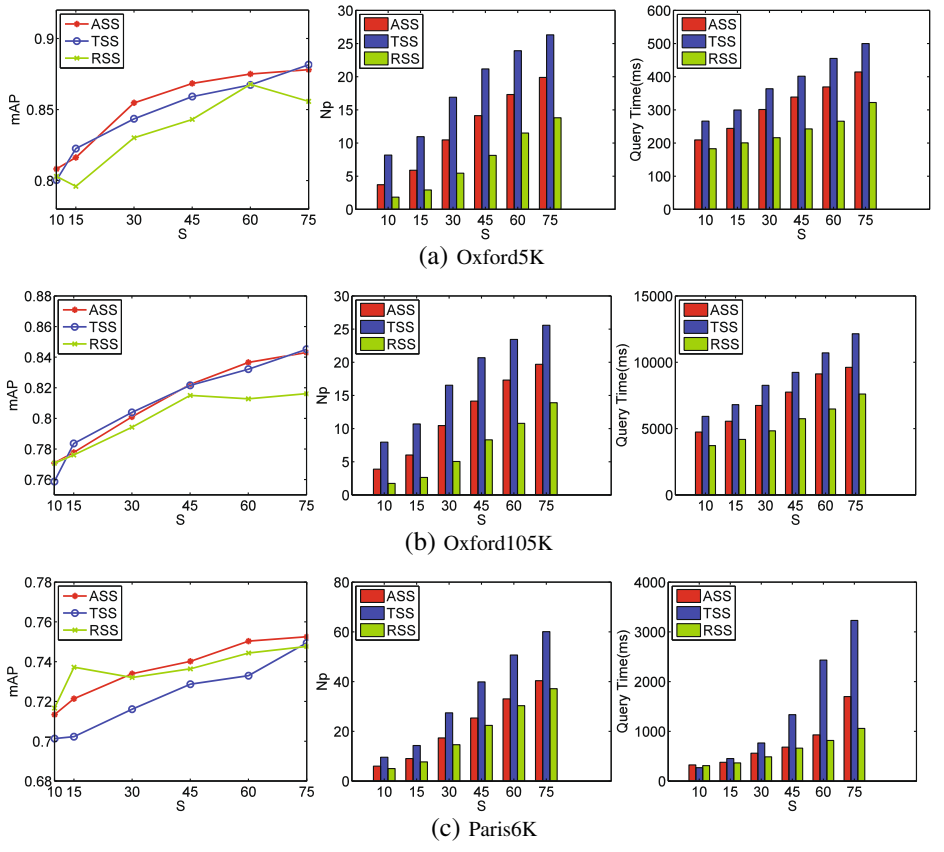


Fig. 6 Mean Average Precision(mAP), number of annotated positives(N_p) and query time t(3.4GHz 8 cores) on different annotation scale (S) in Oxford5K, Oxford105K and Paris6K

annotation scale. Because the samples selected by ASS contain more new visual words about the query. These visual words can help improve the retrieval performance better.

Figure 6. shows the the Mean Average Precision(mAP), the average number of annotated positives(N_p) and the average query time (t) on different scale of annotation(S) with $r = 0.5$. In all the datasets, ASS performs better than TSS when the annotation scale S is small and then they become close with the increase of S. Theoretically, when $S = 200, ASS = TSS = RSS$. We should focus on the performance when S is small, because human annotation costs too much. Both performance of ASS and TSS increase with the rise of S, while RSS does not. When S increases, the expenditure of time of TSS increases a lot, because the number of the visual words increases. ASS achieves a better performance and a faster retrieval speed with less expanded samples than TSS.

In both Oxford datasets the performance of ASS is better than RSS, especially when S is large. $N_p(ASS)$ is much lager than $N_p(RSS)$, which implies the number of positive samples of some queries in the top-200 shortlist of Oxford datasets is small so that RSS can select few positives for query expansion. That is why RSS performs bad in Oxford datasets. In Paris dataset RSS can randomly select positives more easily than in Oxford datasets, so the performance of RSS is better. RSS leads to an uncertain retrieval result, and it rely on a large density of positive samples. ASS is much more stable with RSS in performance on different datasets.

6 Conclusion and further work

In this paper, we addressed the query expansion method under the active learning framework, and we proposed an active sample selection algorithm based on binary relevance classification to explore the most confusing samples with both BoW vector and CNN feature, which may contain useful true positives for query expansion. In our method, we train an binary relevance classifier with the shortlist of basic BoW retrieval as stated in Section 4. And then we pick up the most confusing samples with the classifier for human annotation. The annotated positives will be expanded into new query model for a second query. The proposed method has been evaluated on Oxford5K, Oxford105K and Paris6K datasets. Comparisons with the state-of-the-art methods demonstrate the effectiveness of the proposed method. As shown in the results, our method is able to accurately locate relevant samples and expand the query effectively.

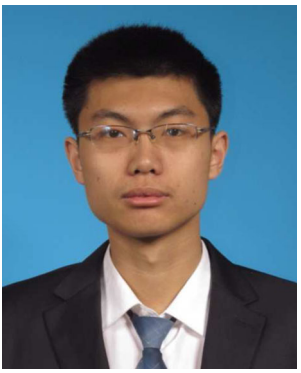
In our experiment, we simply drop the annotated negative samples, which is a waste of information. In the further work we will consider how to take use of the most confusing negative samples for modeling query or filtering other confusers. Furthermore, we also have a lot of work to do with the feature for relevant classification, the method of classifier training and the usage of selected samples.

Acknowledgments This research was supported by the National Natural Science Foundation of China (Grant No.61571269). The authors would like to thank the anonymous reviewers for their valuable comments.

References

1. Arandjelovic R, Zisserman A (2012) Three things everyone should know to improve object retrieval. In: IEEE Conference on Computer Vision and Pattern Recognition, pp 2911–2918
2. Blanchard G, Lee G, Scott C (2010) Semi-supervised novelty detection. J Mach Learn Res 11:2973–3009

3. Chum O, Matas J (2010) Unsupervised discovery of co-occurrence in sparse high dimensional data. In: IEEE Conference on Computer Vision and Pattern Recognition, pp 3416–3423
4. Chum O, Matas J, Kittler J (2003) Locally optimized RANSAC. In: Pattern Recognition, pp 236–243
5. Chum O, Mikulík A, Perdoch M, Matas J (2011) Total recall II: query expansion revisited. In: IEEE Conference on Computer Vision and Pattern Recognition, pp 889–896
6. Chum O, Philbin J, Sivic J, Isard M, Zisserman A (2007) Total recall: Automatic query expansion with a generative feature model for object retrieval. In: International Conference on Computer Vision, pp 1–8
7. du Plessis MC, Niu G, Sugiyama M (2014) Analysis of learning from positive and unlabeled data. In: Ghahramani Z, Welling M, Cortes C, Lawrence N, Weinberger K (eds) NIPS, pp 703–711
8. du Plessis MC, Niu G, Sugiyama M (2015) Convex formulation for learning from positive and unlabeled data. In: International Conference on Machine Learning, pp 1386–1394
9. Jegou H, Douze M, Schmid C (2008) Hamming embedding and weak geometric consistency for large scale image search. In: Computer Vision - European Conference on Computer Vision, pp 304–317
10. Jegou H, Douze M, Schmid C (2009) On the burstiness of visual elements. In: IEEE Conference on Computer Vision and Pattern Recognition, pp 1169–1176
11. Jegou H, Douze M, Schmid C, Pérez P (2010) Aggregating local descriptors into a compact image representation. In: IEEE Conference on Computer Vision and Pattern Recognition, pp 3304–3311
12. Jegou H, Harzallah H, Schmid C (2007) A contextual dissimilarity measure for accurate and efficient image search. In: IEEE Conference on Computer Vision and Pattern Recognition
13. Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. In: NIPS, pp 1106–1114
14. Lebeda K, Matas J, Chum O (2012) Fixing the locally optimized RANSAC. In: BMVC, pp 1–11
15. Lewis DD, Gale WA (1994) A sequential algorithm for training text classifiers. In: SIGIR, pp 3–12
16. Lowe DG (2004) Distinctive image features from scale-invariant keypoints. Int J Comput Vis 60(2):91–110
17. Mikolajczyk K, Schmid C (2004) Scale & affine invariant interest point detectors. Int J Comput Vis 60(1):63–86
18. Mikolajczyk K, Tuytelaars T, Schmid C, Zisserman A, Matas J, Schaffalitzky F, Kadir T, Gool LJV (2005) A comparison of affine region detectors. Int J Comput Vis 65(1-2):43–72
19. Nistér D, Stewénius H (2006) Scalable recognition with a vocabulary tree. In: IEEE Conference on Computer Vision and Pattern Recognition, pp 2161–2168
20. Perdoch M, Chum O, Matas J (2009) Efficient representation of local geometry for large scale object retrieval. In: IEEE Conference on Computer Vision and Pattern Recognition, pp 9–16
21. Perronnin F, Liu Y, Sánchez J, Poirier H (2010) Large-scale image retrieval with compressed fisher vectors. In: IEEE Conference on Computer Vision and Pattern Recognition, pp 3384–3391
22. Philbin J, Chum O, Isard M, Sivic J, Zisserman A (2007) Object retrieval with large vocabularies and fast spatial matching. In: IEEE Conference on Computer Vision and Pattern Recognition
23. Philbin J, Chum O, Isard M, Sivic J, Zisserman A (2008) Lost in quantization: Improving particular object retrieval in large scale image databases. In: IEEE Conference on Computer Vision and Pattern Recognition
24. Sivic J, Zisserman A (2003) Video google: A text retrieval approach to object matching in videos. In: International Conference on Computer Vision, pp 1470–1477
25. Tong S, Chang EY (2001) Support vector machine active learning for image retrieval. In: International Conference on Multimedia, pp 107–118
26. Wan J, Wang D, Hoi SC, Wu P, Zhu J, Zhang Y, Li J (2014) Deep learning for content-based image retrieval: A comprehensive study. In: Proceedings of the ACM International Conference on Multimedia, MM '14, FL, USA, pp 157–166
27. Winder SAJ, Hua G, Brown M (2009) Picking the best DAISY. In: IEEE Conference on Computer Vision and Pattern Recognition, pp 178–185
28. Wu P, Hoi SCH, Xia H, Zhao P, Wang D, Miao C (2013) Online multimodal deep similarity learning with application to image retrieval. In: ACM Multimedia Conference, MM '13, Barcelona, Spain, pp 153–162
29. Zhu C, Jegou H, Satoh S (2013) Query-adaptive asymmetrical dissimilarities for visual object retrieval. In: International Conference on Computer Vision, pp 1705–1712



Xin Zhao is a Ph.D student in School of Software, Tsinghua University. His advisor is Guiguang Ding. He received his bachelor's degree in School of Software, Tsinghua University in 2014. His current research centers on the area of multimedia information retrieval and mining, in particular, visual object retrieval, image caption, content-based multimedia indexing, and image classification.

Guiguang Ding received his Ph.D degree in electronic engineering from the University of Xidian. He is currently an associate professor of School of Software, Tsinghua University. Before joining School of Software in 2006, he worked as a postdoctoral researcher in Automation Department of Tsinghua University. His current research centers on the area of multimedia information retrieval and mining, in particular, visual object classification, automatic semantic annotation, content-based multimedia indexing, and personal recommendation. He has published about 40 research papers in international conferences and journals and applied for 18 Patent Rights in China.