The Thirty-Second AAAI Conference
on Artificial Intelligence (AAAI-18)

Unsupervised Part-Based Weighting Aggregation of Deep Convolutional Features for Image Retrieval

Jian Xu, Cunzhao Shi, Chengzuo Qi, Chunheng Wang,[*] Baihua Xiao
State Key Laboratory of Management and Control for Complex Systems,
Institute of Automation, Chinese Academy of Sciences(CASIA)
University of Chinese Academy of Sciences
{xujian2015, cunzhao.shi, qichengzuo2013, chunheng.wang, baihua.xiao}@ia.ac.cn

Abstract

In this paper, we propose a simple but effective semantic part-based weighting aggregation (PWA) for image retrieval. The proposed PWA utilizes the discriminative filters of deep convolutional layers as part detectors. Moreover, we propose the effective unsupervised strategy to select some part detectors to generate the "probabilistic proposals", which highlight certain discriminative parts of objects and suppress the noise of background. The final global PWA representation could then be acquired by aggregating the regional representations weighted by the selected "probabilistic proposals" corresponding to various semantic content. We conduct comprehensive experiments on four standard datasets and show that our unsupervised PWA outperforms the state-of-the-art unsupervised and supervised aggregation methods.

Introduction

Over the past decades, image retrieval has received sustained attention. The general retrieval framework (Zhou, Li, and Tian 2017) consists of some pivotal modules, i.e., image representation (Husain and Bober 2016; Tolias, Sicre, and Jgou 2016), database indexing (Babenko and Lempitsky 2015b), image scoring (Xie et al. 2015; Zhong, Zhu, and Hoi 2015) and search reranking (Arandjelovic and Zisserman 2012). Image representations derived by aggregating features such as Scale-Invariant Feature Transform (SIFT) (Lowe 2004) and Convolutional Neural Network (CNN) (LeCun et al. 1989) are shown to be effective for image retrieval (Sivic and Zisserman 2003; Jegou et al. 2012; Perronnin and Dance 2007; Gou and Zisserman 2014; Do, Tran, and Cheung 2015; Husain and Bober 2016; Babenko et al. 2014; Razavian et al. 2016; Babenko and Lempitsky 2015a; Tolias, Sicre, and Jgou 2016; Kalantidis, Mellina, and Osindero 2016; Xie et al. 2016; Wei et al. 2017).

Recently, the performance of CNN-based features aggregation methods (Babenko et al. 2014; Razavian et al. 2016; Babenko and Lempitsky 2015a; Tolias, Sicre, and Jgou 2016; Kalantidis, Mellina, and Osindero 2016) rapidly outperforms that of SIFT-based features aggregation methods (Sivic and Zisserman 2003; Jegou et al. 2012; Perronnin and Dance 2007; Perronnin, Nchez, and Mensink 2010; Gou and Zisserman 2014; Do, Tran, and Cheung 2015; Husain and Bober 2016). Some methods (Sharif Razavian et al. 2014; Gong et al. 2014; Babenko et al. 2014) generate the global representation based on fully connected layer features for image retrieval. After that, convolutional features are aggregated to obtain the global representation (Razavian et al. 2016; Babenko and Lempitsky 2015a; Tolias, Sicre, and Jgou 2016; Kalantidis, Mellina, and Osindero 2016; Xie et al. 2016) and achieve better performance. Many recent methods (Arandjelovic et al. 2016; Radenovic, Tolias, and Chum 2016; Gordo et al. 2016a; 2016b) re-train the image representations end-to-end for image retrieval task by collected landmark buildings datasets. The fine-tuning process significantly improves the adaptation ability for the specific task. However, these methods (Arandjelovic et al. 2016; Radenovic, Tolias, and Chum 2016; Gordo et al. 2016a; 2016b) need to collect the labeled training datasets and the performance heavily relies on the collected datasets. The discrepant retrieval objects need different training datasets, for example, the fine-tuned model based on landmarks is not suitable for logo retrieval.

Previous aggregation methods ignore the discriminative information from the object parts. The part-based information is utilized for fine-gained categorization (Zhang et al. 2016a; Xiao et al. 2015; Simon and Rodner 2015; Zhang et al. 2016b; He and Peng 2017) and the part-based representation provides the state-of-the-art performance. Zhang et al. (Zhang et al. 2016a) pick some distinctive filters which respond to specific patterns significantly and consistently to learn a set of part detectors. Then, they conditionally encode the deep filter responses into the final representation based on Fisher vector (Perronnin and Dance 2007). In recent work (Zhang et al. 2016b), the part-based image representation is generated by aggregating selected parts on several different scales. The recent work (He and Peng 2017) applies spatial constraints to select part proposals which are generated by selective search (Uijlings et al. 2013). Different with these methods, the selected parts proposals ("probabilistic proposals") in our algorithm are not constrained to rectangular box but erose shape.

Some recent works (He et al. 2014; Zhang et al. 2016a; Zeiler and Fergus 2014) analyze the meaning of feature maps of CNN. Zeiler et al. (Zeiler and Fergus 2014) show that some input patterns stimulate the special channels of

[*]Corresponding author

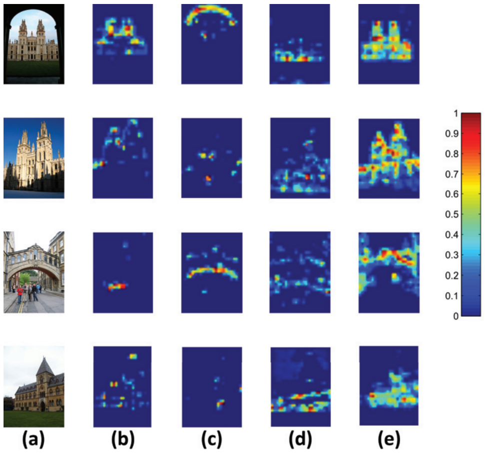

Figure 1: Visualization of "probabilistic proposals". (a) Some images in Oxford5K (Philbin et al. 2007). (b)-(e) The various channels of feature maps in *pool*5 layer from pre-trained VGG16 (Simonyan and Zisserman 2015). Each channel of feature maps is activated (warm) by different parts or patterns of objects and some discriminative channels can work as "probabilistic proposals".

feature maps of the latter convolutional layers. He et al. visualize the feature maps generated by some filters of the $conv_5$ layer from SPP-net (He et al. 2014) and show that the filters of deep convolutional layers are activated by specific semantic content and some distinctive filters can work as part detectors. The various channels of convolutional feature maps can represent the pixel-level label mask of different categories in Fully Convolutional Network (FCN) (Long, Shelhamer, and Darrell 2015). Instance-aware semantic segmentation (Dai, He, and Sun 2016; Li et al. 2017) employs the different channels of shared convolutional layers to detect and segment the various object instance jointly. Mask R-CNN (He et al. 2017) demonstrates that the erose proposals perform better than the rectangular regions on object detection task. Inspired by above works, we employ some selected discriminative filters of deep convolutional layers as the part detectors to generate erose "probabilistic proposals", which correspond to fixed semantic content implicitly.

In this paper, we define the special channel of normalized feature maps as "probabilistic proposal". The "probabilistic proposal" encodes the spatial layout of input object's parts corresponding to various semantic content, and represents the probability of pixels belonging to fixed semantic. To further understand the meanings and characteristics of the "probabilistic proposals", we visualize some images and corresponding typical "probabilistic proposals" in Fig. 1. We select some images in Oxford5K (Philbin et al. 2007) as shown in Fig. 1 (a). In Fig. 1 (b)-(e), we visualize some discriminative channels of feature maps which work as the "probabilistic proposals" for the selected images. Each channel of feature maps is activated (warm) by special parts or patterns corresponding to fixed semantic content and the background is suppressed (cold). For example, the 220th feature map (Fig. 1 (b)) of *pool*5 layers from VGG16 (Simonyan and Zisserman 2015) is most activated by the sharp shape; the 478th feature map (Fig. 1 (c)) is most activated by the arc shape; the 483th feature map (Fig. 1 (d)) is most activated by the bottom of buildings; the 360th feature map (Fig. 1 (e)) is most activated by the body of buildings. We can see that different filters of deep convolutional layers are sensitive to different shapes or semantic, and they highlight different parts and patterns of objects. Some special parts of object are discriminative, for example, the 220th feature maps highlight the spire of buildings. Therefore, filters of deep convolutional layers can work as part detectors to pick special patterns corresponding to fixed semantic content. We select the discriminative filters of deep convolutional layers as the part detectors to generate erose "probabilistic proposals, which are related to different semantic content.

Inspired by the characteristics of feature maps, in this paper we propose a novel and simple way of creating powerful image representation via part-based aggregation. Our unsupervised part-based weighting aggregation (PWA) method significantly outperforms the state-of-the-art unsupervised aggregation methods (Razavian et al. 2016; Babenko and Lempitsky 2015a; Tolias, Sicre, and Jgou 2016; Kalantidis, Mellina, and Osindero 2016) and supervised methods (Radenovic, Tolias, and Chum 2016; Gordo et al. 2016a; 2016b) on four standard retrieval datasets.

The main contributions of this paper can be summarized as follows:

"Probabilistic proposal" We select some discriminative part detectors by succinct unsupervised strategy to generate the "probabilistic proposals" corresponding to special semantic content. Different with previous methods, the selected "probabilistic proposals" are not constrained to rectangular box and represent the confidence degree of fixed semantic. To the best of our knowledge, this paper is the first work to select the erose "probabilistic proposals" for image retrieval, and the selected "probabilistic proposals" corresponding to special semantic content are tactfully employed to generate high-dimensional representation which contains discriminative semantic information.

Part-based weighting aggregation We aggregate the convolutional features weighted by selected "probabilistic proposals" and concatenate the regional representations as global PWA representation. Because selected "probabilistic proposals" corresponds to fixed semantic but not fixed position, the selected regional representations can be concatenated as the global PWA representation. Concatenation as the global representation preserves more discrimination than summing regional representations.

Aggregation based on "probabilistic proposals"

The diagram of the proposed method is shown in Fig. 6. Based on the dataset, we pick the discriminative part detectors to generate the "probabilistic proposals" by the unsupervised strategy in the off-line stage. Each "probabilistic proposal" corresponds to fixed semantic content implic-

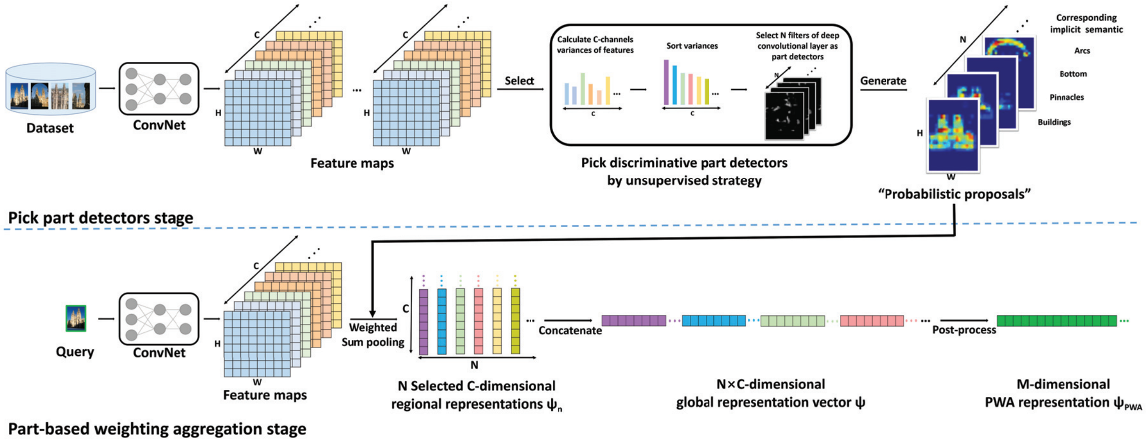
Dataset
ConvNet
C
H
W
Feature maps
Select
Calculate C-channels variances of features
Sort variances
Select N filters of deep convolutional layer as part detectors
N
c
Pick discriminative part detectors by unsupervised strategy
Generate
Corresponding implicit semantic
Arcs
Bottom
Pinnacles
Buildings
"Probabilistic proposals"
Pick part detectors stage
Query
ConvNet
Feature maps
Weighted Sum pooling
N Selected C-dimensional regional representations ψ_n
Concatenate
N×C-dimensional global representation vector ψ
Post-process
M-dimensional PWA representation ψ_PWA
Part-based weighting aggregation stage

Figure 2: Flow chart of our part-based weighting aggregation (PWA) method. We pick the discriminative part detectors to generate the "probabilistic proposals" by the unsupervised strategy in the first off-line stage. Each "probabilistic proposal" corresponds to fixed semantic content implicitly, such as pinnacles, arcs and bottom of buildings. In the aggregation stage, we employ the selected N "probabilistic proposals" to weight and aggregate the feature maps as C-dimensional regional representations, and concatenate N regional representations as the final global PWA representation.

itly, such as pinnacles, arcs and bottom of buildings. In the aggregation stage, we employ the selected N "probabilistic proposals" to weight and aggregate the feature maps as C-dimensional regional representations. Finally, we concatenate N regional representations corresponding to special sematic content as the final global PWA representation.

In this section, we analyse the characteristics of the filters of deep convolutional layers which can be interpreted as part detectors. We propose the unsupervised strategy to select discriminative part detectors to generate "probabilistic proposals". Based on the selected "probabilistic proposals" corresponding to special semantic content, we propose a novel and effective PWA aggregation method for image retrieval.

We extract features f from deep convolutional layers by passing an image I through a pre-trained or fine-tuned deep network, which consist of C channels feature maps each with height H and width W. Finally, the input image I is represented by the aggregated $N \times C$-dimensional vector that are weighted by the N selected part detectors.

"Probabilistic proposals"

Selection of part detectors Because the responses with large variances are significantly different among the various objects, the channels of feature maps with large variances are more discriminative. Therefore, we select part detectors according to variances based on dataset.

We first calculate the C-channels variances $V = \{v_1, v_2, ..., v_c, ..., v_C\}$ of the C-dimensional vectors g_i $(i = 1, 2, ..., D)$ computed by sum pooling the $C \times W \times H$-dimensional deep convolutional features f_i of image i.

$$V = \frac{1}{D} \sum_{i=1}^{D} (g_{_i} - \bar{g})^2 \tag{1}$$

where D is the number of database images. $\bar{g} = \frac{1}{D} \sum\limits_{i=1}^{D} g_i$ is the average vector of feature vectors g_i $(i = 1, 2, ..., D)$.

$$g_i = \sum_{x=1}^{W} \sum_{y=1}^{H} f_i(x, y) \tag{2}$$

Then we sort the variances $\{v_1, v_2, ..., v_C\}$ of C channels. We select the discriminative deep convolutional layers filters corresponding to large variances as the part detectors. We also observe the filters with large variances to be more discriminative by the following experiment. We performed retrieval by PWA but we select (1) 30% random part detectors (2) 30% part detectors with the largest variance. The mAP score for the Oxford5k dataset (Philbin et al. 2007) for (1) is only 0.775±0.006, which is much small than mAP for (2), 0.790. This verifies that feature maps with large variances are much more discriminative than random feature maps. Moreover, our simple unsupervised selection method not only boosts the performance but also reduces the computational complexity of PWA representation.

Effects of "probabilistic proposals" The special channels of feature maps generated by selected part detectors can work as the "probabilistic proposals" corresponding to fixed semantic content. To investigate the effects of "probabilistic proposals" in detail, we compare the 512-dimensional representation computed by sum pooling with the representation weighted by the discriminative "probabilistic proposals" in Fig. 3. As shown in Fig. 3, the selected "probabilistic proposal" generated by 220th part detector suppresses the noise of background and activates the sharp shape. Weighted by the selected "probabilistic proposal", the values of feature maps that are activated by background (such as (a) 507th and (b) 155th) are smaller. However, the values of the representation corresponding to similar semantic content to the

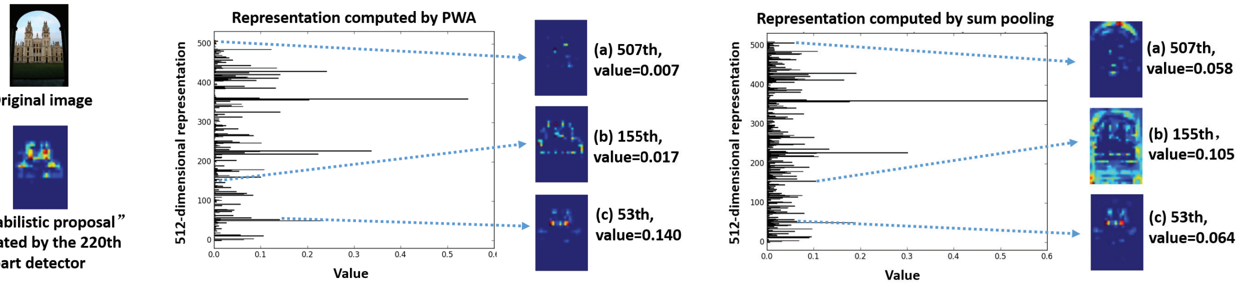

Figure 3: The comparison of the 512-dimensional representations computed by PWA and sum pooling. Weighted by the selected “probabilistic proposal”, values of the feature map’s channels activated by background (such as (a) 507th and (b) 155th) are reduced. However, values of the representation corresponding to similar patterns to the selected “probabilistic proposal” (such as (c) 53th) still keep large. The selected “probabilistic proposal” suppresses (cold) the noise of background and highlights (warm) the special semantic content.

selected “probabilistic proposal” (such as (c) 53th) still keep large. As a result, the representations weighted by the discriminative “probabilistic proposals” are more discriminative and robust.

Overall, the discriminative filters of latter convolutional layers are interpreted as part detectors to generate the “probabilistic proposals”. The selected “probabilistic proposals” suppress the noise of background and highlight the discriminative parts and patterns of objects. We make use of the selected “probabilistic proposals” to weight the activations of convolutional layers and generate the regional representations. Because each filter of deep convolutional layers activates special pattern, the various selected part detectors can be employed to generate the erose proposals corresponding to special semantic content. Each proposal corresponds to a fixed semantic pattern implicitly. The erose “probabilistic proposals” maintain the explicit $W \times H$ object spatial layout which can be addressed naturally by the pixel-to-pixel correspondence provided by convolutions. Different with R-MAC (Tolias, Sicre, and Jgou 2016), the “probabilistic proposals” corresponds to fixed semantic content rather than fixed position. Our “probabilistic proposals” are not constrained to box and represent the probability of pixels belong to fixed semantic content. Although the “probabilistic proposals” corresponding to the part detectors selected by unsupervised strategy do not explicitly describe the semantic, they implicitly represent discriminative semantic content, such as pinnacles, arcs and bottom of buildings. Therefore we can concatenate the selected regional representations weighted by special semantic “probabilistic proposals” as the final global representation. CroW (Kalantidis, Mellina, and Osindero 2016), InterActive (Xie et al. 2016) and PWA can be interpreted as spatial-weighted representations. InterActive (Xie et al. 2016) is much more generalized in the aspect of spatial-weighted, which integrates high-level visual context with low-level neuron responses by back-propagation. Compared to CroW (Kalantidis, Mellina, and Osindero 2016) and InterActive (Xie et al. 2016) that sum the spatial-weighted representations, we independently employ the selected part detectors to extract the regional representations corresponding to special semantic content and concatenate them as final PWA representation. The concatenation of regional representations preserves more discriminative information than summation in R-MAC (Tolias, Sicre, and Jgou 2016), CroW (Kalantidis, Mellina, and Osindero 2016) and InterActive (Xie et al. 2016).

PWA design

In this section, we describe the PWA method in detail. We aggregate the feature maps weighted by the selected “probabilistic proposals” and concatenate the regional representations as global PWA representation. We reduce the dimensionality of high-dimensional PWA representation by unsupervised method (PCA) in post-processing.

Weighted by selected “probabilistic proposals” The construction of the PWA representation starts with the weighted sum pooling of the $C \times W \times H$-dimensional deep convolutional features f of image I with height H and width W:

$$\psi_{\mathrm{n}}(I) = \sum_{x=1}^{W} \sum_{y=1}^{H} w_n(x,y) f(x,y) \tag{3}$$

The coefficients w_n are the normalized weights as follows, which depend on the activation values $v_n(x,y)$ in position (x,y) of the selected “probabilistic proposal” generated by part detector n:

$$w_n(x,y) = \left(\frac{v_n(x,y)}{\left(\sum\limits_{x=1}^{W} \sum\limits_{y=1}^{H} v_n(x,y)^{\alpha} \right)^{1/\alpha}} \right)^{1/\beta} \tag{4}$$

where α and β are parameters of power normalization and power-scaling respectively.

Concatenation N selected C-dimensional regional representations $\psi_n(I)$ are obtained from weighted sum pooling process. We get the global $N \times C$-dimensional representation vector $\psi(I)$ by concatenating selected regional representations:

$$\psi(I) = [\psi_1, \psi_2, \cdots \psi_N] \tag{5}$$

where we select the N part detectors depending on the discrimination of them. The selection based on the values of the variances of different C channels of feature maps both provides boost in performance and enhances the computation efficiency.

Post-processing We perform l_2-normalization, PCA compression and whitening on the global representation $\psi(I)$ subsequently and obtain the final M-dimensional representation $\psi_{PWA}(I)$:

$$\psi_{PWA}(I) = diag(\sigma_1, \sigma_2, \cdots, \sigma_M)^{-1} V \frac{\psi(I)}{\|\psi(I)\|_2} \tag{6}$$

where V is the $M \times N$ PCA-matrix, M is the number of the retained dimensionality, and $\sigma_1, \sigma_2, \cdots, \sigma_M$ are the associated singular values.

Experiments

Datasets

We evaluate the performance of PWA and other aggregation algorithms on four standard datasets (Oxford5k, Paris6k, Oxford105k and Paris106k) for image retrieval.

Oxford5k (Philbin et al. 2007) and Paris6k (Philbin et al. 2008) datasets contain photographs collected from Flickr associated with Oxford and Paris landmarks respectively. The performance is measured using mean average precision (mAP) over the 55 queries annotated manually. Oxford105k and Paris106k contain the additional 10,000 distractor images from Flicker (Philbin et al. 2007).

Implementation details

We extract deep convolutional features using the pre-trained VGG16 (Simonyan and Zisserman 2015) and fine-tuned ResNet101 from the work (Gordo et al. 2016b). In the experiments, Caffe (Jia et al. 2014) package for CNNs is used. For VGG16 model, we extract convolutional feature maps from the $pool5$ layer and the number of channels is C=512. For ResNet-101 model, we extract convolutional feature maps from the $res5c\_relu$ layer and the number of channels is C=2048. Regarding image size, we keep the original size of the images except for the very large images which are resized to the half size. The parameters for power normalization and power-scaling are set as $\alpha = 2$ and $\beta = 2$, throughout our experiments.

We evaluate the mean average precision (mAP) over the cropped query. For fair comparison with the related retrieval methods, we learn the PCA and whitening parameters on Oxford5k when testing on Paris6k and vice versa.

Impact of the parameters

The main parameters are the numbers of the selected part detectors and the dimensionality of final representations $\psi_{PWA}(I)$.

Select part detectors We employ the discriminative filters of deep convolutional layers as part detectors to generate "probabilistic proposals". The discriminative part detectors are selected according to the variances of C channels of feature maps. We also aggregate the responses of convolutional layers based on all the C part detectors as the baseline. We show the results of selecting the first N part detectors with the largest variance in Table 1. In this experiment, the final representation $\psi_{PWA}(I)$ is reduced into 4096 dimensionality by PCA.

Table 1: Performance of different number of selected part detectors (N). We aggregate the responses of convolutional layers by all the C=512 part detectors as the baseline. Note, the final representation $\psi_{PWA}(I)$ is reduced into 4096 dimensionality by PCA.

	Datasets	
N	Oxford5k	Paris6k
512	78.5	85.4
450	78.7	85.7
350	79.0	85.9
250	78.7	86.0
150	79.0	85.4
50	78.2	86.1
25	**79.1**	**86.1**
10	77.7	83.8

The results show that our PWA representation is not heavily relied on the number of selected part detectors. Selecting a small number of part detectors (e.g., N=25), we still achieve good performance. The selection strategy boosts above 0.6% mAP than baseline and reduces the computational cost to $1/20$ of the baseline. The results demonstrate that our straightforward unsupervised selection strategy is effective.

Table 2: Performance of varying dimensionality (M), into which the final representation is reduced. The representation is reduced by PCA and whitening. Note, we select 25 part detectors to aggregate the convolutional features.

	Datasets	
M	Oxford5k	Paris6k
128	64.5	76.9
256	68.7	79.6
512	72.0	82.3
1024	75.3	84.2
2048	78.2	85.4
4096	**79.1**	**86.1**

Dimensionality reduction In order to get shorter representations, we compress the $N \times C$-dimensional aggregated representation $\psi(I)$ by PCA and whitening process. Table 2 reports the performance of representations with varying dimensionality, M=128 to 4096. We do not reduce the final representation into higher dimensionality because of the limited number of images in Oxford5k and Paris6k datasets. We select N=25 part detectors to aggregate the convolutional features in this experiment.

Table 3: Accuracy comparison with the state-of-the-art unsupervised methods. We compare our PWA+QE with other methods followed by query expansion at the bottom of table. Part-based weighting aggregation (PWA) consistently outperforms the state-of-the-art unsupervised aggregation methods.

		Datasets			
Method	**Dimensionality**	Oxford5k	Paris6k	Oxford105k	Paris106k
Tri-embedding (Gou and Zisserman 2014)	8k	67.6	—	61.1	—
FAemb (Do, Tran, and Cheung 2015)	16k	70.9	—	—	—
RVD-W (Husain and Bober 2016)	16k	68.9	—	66.0	—
Razavian et al. (Razavian et al. 2016)	512	46.2	67.4	—	—
Neural Codes (Babenko et al. 2014)	512	43.5	—	39.2	—
SPoC (Babenko and Lempitsky 2015a)	256	53.1	—	50.1	—
InterActive (Xie et al. 2016)	512	65.6	79.2	—	—
R-MAC (Tolias, Sicre, and Jgou 2016)	512	66.9	83.0	61.6	75.7
CroW (Kalantidis, Mellina, and Osindero 2016)	512	70.8	79.7	65.3	72.2
Previous state-of-the-art		70.8	83.0	65.3	75.7
PWA	512	**72.0**	82.3	**66.2**	**75.8**
PWA	1024	75.3	84.2	69.3	78.2
PWA	2048	78.2	85.4	71.1	79.7
PWA	4096	**79.1**	**86.1**	**73.6**	**80.4**
CroW+QE (Kalantidis, Mellina, and Osindero 2016)	512	74.9	84.8	70.6	79.4
R-MAC+AML+QE (Tolias, Sicre, and Jgou 2016)	512	77.3	86.5	73.2	79.8
DSM (Zhong, Zhu, and Hoi 2015)	—	95.0	91.5	93.2	—
PWA+QE	512	74.8	86.0	72.5	80.7
PWA+QE	1024	77.9	87.8	76.7	82.8
PWA+QE	2048	80.7	88.7	79.3	83.9
PWA+QE	4096	**81.7**	**89.2**	**80.6**	**84.7**

The results show that the performance boosts gradually with the increase of dimensionality and the best performance is achieved at 4096 dimensionality. We get the consistent conclusion with other methods, the compression leads to the loss of discriminative information and performance degradation. The previous works (Babenko and Lempitsky 2015a; Tolias, Sicre, and Jgou 2016; Kalantidis, Mellina, and Osindero 2016) aggregate convolutional features as compressed representations with dimensionality under 512, but our PWA representation has more choice of dimensionality. Compared with (Babenko and Lempitsky 2015a; Tolias, Sicre, and Jgou 2016; Kalantidis, Mellina, and Osindero 2016), our PWA methods can generate representations with both low and high dimensionality and achieve better performance on most datasets. The dimensionality of PWA representation can be chosen according to the tradeoff between performance and efficiency on different tasks.

Comparison with the state-of-the-art

Unsupervised methods In the first part of Table 3, we compare our PWA method using pre-trained VGG16 (Simonyan and Zisserman 2015) with the state-of-the-art unsupervised methods, which employ global representations of images. Our PWA representation significantly outperform them on all four standard retrieval datasets. In particular, the gain is more than 8.3% in mAP on Oxford5k and Oxford105k datasets. The results demonstrate that our PWA representation weighted by the selected “probabilistic proposals” is effective and discriminative for image retrieval. Our 512-dimensional PWA representation is comparable with the previous state-of-the-art, and its results are only lower than R-MAC (Tolias, Sicre, and Jgou 2016) on Paris6k. The PWA representation with higher dimensionality (such as 1024, 2048 and 4096) consistently outperform all of them on all datasets.

We compare other methods that contain query expansion (QE) and spatial verification stages with our approach in the second part of Table 3. In the experiments, we use average query expansion (QE) (Chum et al. 2007) computed by the top 10 query results. Our PWA+QE method performs better than the related works (Tolias, Sicre, and Jgou 2016; Kalantidis, Mellina, and Osindero 2016) on all datasets. Although the approximate max pooling localization (AML) process in R-MAC (Tolias, Sicre, and Jgou 2016) requires a costly verification stage and the extra memory storage, our PWA+QE still achieves better performance than R-MAC+AML+QE. DSM (Zhong, Zhu, and Hoi 2015) uses handcrafted features(SIFT) and achieves better performance by employing additional time-consuming reranking processes ,i.e., spatial verification in tf-idf and k-NN reranking.

Supervised methods with end-to-end training We also compare our method with the current state-of-the-art supervised methods containing end-to-end training process (Arandjelovic et al. 2016; Radenovic, Tolias, and Chum 2016; Gordo et al. 2016b)) in Table 4. In order to compare

Table 4: Accuracy comparison with the state-of-the-art supervised methods. Employing the convolutional layer features of fine-tuned network (Gordo et al. 2016b), we achieve the comparable performance with the state-of-the-art methods with end-to-end supervised training.

		Datasets			
Method	**Dimensionality**	Oxford5k	Paris6k	Oxford105k	Paris106k
NetVLAD (Arandjelovic et al. 2016)	4096	71.6	79.7	—	—
CNNBoW (Radenovic, Tolias, and Chum 2016)	512	79.7	83.8	73.9	76.4
DeepRepresentation (Gordo et al. 2016b)	2048	86.1	94.5	82.8	90.6
Previous state-of-the-art		86.1	94.5	82.8	90.6
PWA	2048	**87.8**	**94.9**	**82.8**	**91.0**

with them, we employ convolutional layers features of fine-tuned ResNet101 from the work (Gordo et al. 2016b). Because these methods (Radenovic, Tolias, and Chum 2016; Gordo et al. 2016b) map the final representation by the supervised methods for similarity evaluation, we also map the PWA representations for comparison purposes. In order to keep consistently unsupervised, we utilize the unsupervised IME layer (Xu et al. 2017) to map our PWA representations for similarity evaluation.

The results show that our unsupervised PWA representation outperforms the state-of-the-art supervised methods (Arandjelovic et al. 2016; Radenovic, Tolias, and Chum 2016; Gordo et al. 2016b) on all datasets. Furthermore, the effectiveness of the supervised methods is heavily relied on the collected training set. However, our unsupervised PWA method can make better use of the convolutional features extracted from both pre-trained and fine-tuned CNN model to represent the images and does not need the further supervised re-training. Considering the fact that the annotated training dataset is difficult to collect, it is impractical to fine-tune the model for each discrepant task respectively. Our unsupervised PWA method is very suitable for this condition. Our PWA method retains more discriminative information of the retrieval object parts and significantly suppress the noise of background, and better utilizes the convolutional features extracted from both pre-trained and fine-tuned CNN models.

Conclusion

In this paper, we propose a novel PWA method for image retrieval. The key characteristic of our method is that it employs discriminative part detectors selected by unsupervised strategy to generate "probabilistic proposals". Based on the selected "probabilistic proposals" corresponding to special semantic content implicitly, we weight and aggregate the deep convolutional features extracted from pre-trained or fine-tuned CNN models. Due to the selected "probabilistic proposals" corresponding to fixed semantic content but not fixed position, we concatenate the regional representations as global PWA representation. The results show that our PWA representation suppress the noise of background and highlight the discriminative parts and patterns of retrieval objects.

Experiments on four standard retrieval datasets demonstrate that our unsupervised approach outperforms the previous state-of-the-art unsupervised and supervised aggregation methods. It is worth noting that our unsupervised PWA method is very suitable and effective for the situation where the annotated training dataset is difficult to collect.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grant 61531019, Grant 61601462, and Grant 71621002.

References

Arandjelovic, R., and Zisserman, A. 2012. Three things everyone should know to improve object retrieval. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2911–2918. IEEE.

Arandjelovic, R.; Gronat, P.; Torii, A.; Pajdla, T.; and Sivic, J. 2016. Netvlad: Cnn architecture for weakly supervised place recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 5297–5307.

Babenko, A., and Lempitsky, V. 2015a. Aggregating local deep features for image retrieval. In *IEEE international conference on computer vision*, 1269–1277.

Babenko, A., and Lempitsky, V. 2015b. The inverted multi-index. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 37(6):1247–1260.

Babenko, A.; Slesarev, A.; Chigorin, A.; and Lempitsky, V. 2014. Neural codes for image retrieval. In *European conference on computer vision*, 584–599. Springer.

Chum, O.; Philbin, J.; Sivic, J.; Isard, M.; and Zisserman, A. 2007. Total recall: Automatic query expansion with a generative feature model for object retrieval. In *IEEE International Conference on Computer Vision*, 1–8. IEEE.

Dai, J.; He, K.; and Sun, J. 2016. Instance-aware semantic segmentation via multi-task network cascades. In *IEEE Conference on Computer Vision and Pattern Recognition*, 3150–3158.

Do, T.-T.; Tran, Q. D.; and Cheung, N.-M. 2015. Faemb: a function approximation-based embedding method for image retrieval. In *IEEE Conference on Computer Vision and Pattern Recognition*, 3556–3564.

Gong, Y.; Wang, L.; Guo, R.; and Lazebnik, S. 2014. Multi-scale orderless pooling of deep convolutional activation features. In *European conference on computer vision*, 392–407. Springer.

Gordo, A.; Almazan, J.; Revaud, J.; and Larlus, D. 2016a. Deep image retrieval: Learning global representations for image search. In *European Conference on Computer Vision*, 241–257. Springer.

Gordo, A.; Almazan, J.; Revaud, J.; and Larlus, D. 2016b. End-to-end learning of deep visual representations for image retrieval. *International Journal of Computer Vision* 1–18.

Gou, H., and Zisserman, A. 2014. Triangulation embedding and democratic aggregation for image search. In *Computer Vision and Pattern Recognition*, 3310–3317.

He, X., and Peng, Y. 2017. Weakly supervised learning of part selection model with spatial constraints for fine-grained image classification. In *AAAI*, 4075–4081.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2014. Spatial pyramid pooling in deep convolutional networks for visual recognition. In *European Conference on Computer Vision*, 346–361. Springer.

He, K.; Gkioxari, G.; Dollar, P.; and Girshick, R. 2017. Mask r-cnn. *IEEE International Conference on Computer Vision*.

Husain, S. S., and Bober, M. 2016. Improving large-scale image retrieval through robust aggregation of local descriptors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Jegou, H.; Perronnin, F.; Douze, M.; Sanchez, J.; Perez, P.; and Schmid, C. 2012. Aggregating local image descriptors into compact codes. *IEEE transactions on pattern analysis and machine intelligence* 34(9):1704–1716.

Jia, Y.; Shelhamer, E.; Donahue, J.; Karayev, S.; Long, J.; Girshick, R.; Guadarrama, S.; and Darrell, T. 2014. Caffe: Convolutional architecture for fast feature embedding. In *ACM international conference on Multimedia*, 675–678. ACM.

Kalantidis, Y.; Mellina, C.; and Osindero, S. 2016. Cross-dimensional weighting for aggregated deep convolutional features. In *European Conference on Computer Vision*, 685–701. Springer.

LeCun, Y.; Boser, B.; Denker, J. S.; Henderson, D.; Howard, R. E.; Hubbard, W.; and Jackel, L. D. 1989. Backpropagation applied to handwritten zip code recognition. *Neural computation* 1(4):541–551.

Li, Y.; Qi, H.; Dai, J.; Ji, X.; and Wei, Y. 2017. Fully convolutional instance-aware semantic segmentation.

Long, J.; Shelhamer, E.; and Darrell, T. 2015. Fully convolutional networks for semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 3431–3440.

Lowe, D. G. 2004. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision* 60(60):91–110.

Perronnin, F., and Dance, C. 2007. Fisher kernels on visual vocabularies for image categorization. In *IEEE Conference on Computer Vision and Pattern Recognition*, 1–8.

Perronnin, F.; Nchez, J.; and Mensink, T. 2010. Improving the fisher kernel for large-scale image classification. In *European Conference on Computer Vision*, 143–156.

Philbin, J.; Chum, O.; Isard, M.; Sivic, J.; and Zisserman, A. 2007. Object retrieval with large vocabularies and fast spatial matching. In *IEEE Conference on Computer Vision and Pattern Recognition*, 1–8. IEEE.

Philbin, J.; Chum, O.; Isard, M.; Sivic, J.; and Zisserman, A. 2008. Lost in quantization: Improving particular object retrieval in large scale image databases. In *IEEE Conference on Computer Vision and Pattern Recognition*, 1–8. IEEE.

Radenovic, F.; Tolias, G.; and Chum, O. 2016. Cnn image retrieval learns from bow: Unsupervised fine-tuning with hard examples. In *European Conference on Computer Vision*, 3–20. Springer.

Razavian, A. S.; Sullivan, J.; Carlsson, S.; and Maki, A. 2016. Visual instance retrieval with deep convolutional networks. *ITE Transactions on Media Technology and Applications* 4(3):251–258.

Sharif Razavian, A.; Azizpour, H.; Sullivan, J.; and Carlsson, S. 2014. Cnn features off-the-shelf: an astounding baseline for recognition. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 806–813.

Simon, M., and Rodner, E. 2015. Neural activation constellations: Unsupervised part model discovery with convolutional networks. In *IEEE International Conference on Computer Vision*, 1143–1151.

Simonyan, K., and Zisserman, A. 2015. Very deep convolutional networks for large-scale image recognition. *ICLR*.

Sivic, J., and Zisserman, A. 2003. Video google: A text retrieval approach to object matching in videos. In *IEEE International Conference on Computer Vision*, 1470.

Tolias, G.; Sicre, R.; and Jgou, H. 2016. Particular object retrieval with integral max-pooling of cnn activations. *ICLR*.

Uijlings, J. R.; Van De Sande, K. E.; Gevers, T.; and Smeulders, A. W. 2013. Selective search for object recognition. *International journal of computer vision* 104(2):154–171.

Wei, X.-S.; Luo, J.-H.; Wu, J.; and Zhou, Z.-H. 2017. Selective convolutional descriptor aggregation for fine-grained image retrieval. *IEEE Transactions on Image Processing* 26(6):2868–2881.

Xiao, T.; Xu, Y.; Yang, K.; Zhang, J.; Peng, Y.; and Zhang, Z. 2015. The application of two-level attention models in deep convolutional neural network for fine-grained image classification. In *IEEE Conference on Computer Vision and Pattern Recognition*, 842–850.

Xie, L.; Hong, R.; Zhang, B.; and Tian, Q. 2015. Image classification and retrieval are one. In *ACM on International Conference on Multimedia Retrieval*, 3–10. ACM.

Xie, L.; Zheng, L.; Wang, J.; Yuille, A. L.; and Tian, Q. 2016. Interactive: Inter-layer activeness propagation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 270–279.

Xu, J.; Wang, C.; Qi, C.; Shi, C.; and Xiao, B. 2017. Iterative manifold embedding layer learned by incomplete data for large-scale image retrieval. *arXiv preprint arXiv:1707.09862*.

Zeiler, M. D., and Fergus, R. 2014. Visualizing and understanding convolutional networks. In *European conference on computer vision*, 818–833. Springer.

Zhang, X.; Xiong, H.; Zhou, W.; Lin, W.; and Tian, Q. 2016a. Picking deep filter responses for fine-grained image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 1134–1142.

Zhang, Y.; Wei, X.-S.; Wu, J.; Cai, J.; Lu, J.; Nguyen, V.-A.; and Do, M. N. 2016b. Weakly supervised fine-grained categorization with part-based image representation. *IEEE Transactions on Image Processing* 25(4):1713–1725.

Zhong, Z.; Zhu, J.; and Hoi, S. C. 2015. Fast object retrieval using direct spatial matching. *IEEE Transactions on Multimedia* 17(8):1391–1397.

Zhou, W.; Li, H.; and Tian, Q. 2017. Recent advance in content-based image retrieval: A literature survey. *arXiv preprint arXiv:1706.06064*.