Journal of New Media JNM, vol.1, no.2, pp.87-99, 2019

Instance Retrieval Using Region of Interest Based CNN Features

Jingcheng Chen[1], Zhili Zhou[1, 2, *], Zhaoqing Pan[1] and Ching-nung Yang[3]

Abstract: Recently, image representations derived by convolutional neural networks (CNN) have achieved promising performance for instance retrieval, and they outperform the traditional hand-crafted image features. However, most of existing CNN-based features are proposed to describe the entire images, and thus they are less robust to background clutter. This paper proposes a region of interest (RoI)-based deep convolutional representation for instance retrieval. It first detects the region of interests (RoIs) from an image, and then extracts a set of RoI-based CNN features from the fully-connected layer of CNN. The proposed RoI-based CNN feature describes the patterns of the detected RoIs, so that the visual matching can be implemented at image region-level to effectively identify target objects from cluttered backgrounds. Moreover, we test the performance of the proposed RoI-based CNN feature, when it is extracted from different convolutional layers or fully-connected layers. Also, we compare the performance of RoI-based CNN feature with those of the state-of-the-art CNN features on two instance retrieval benchmarks. Experimental results show that the proposed RoI-based CNN feature provides superior performance than the state-of-the-art CNN features for in-stance retrieval.

Keywords: Image retrieval, instance retrieval, RoI, CNN, convolutional layer, convolutional feature maps.

1 Introduction

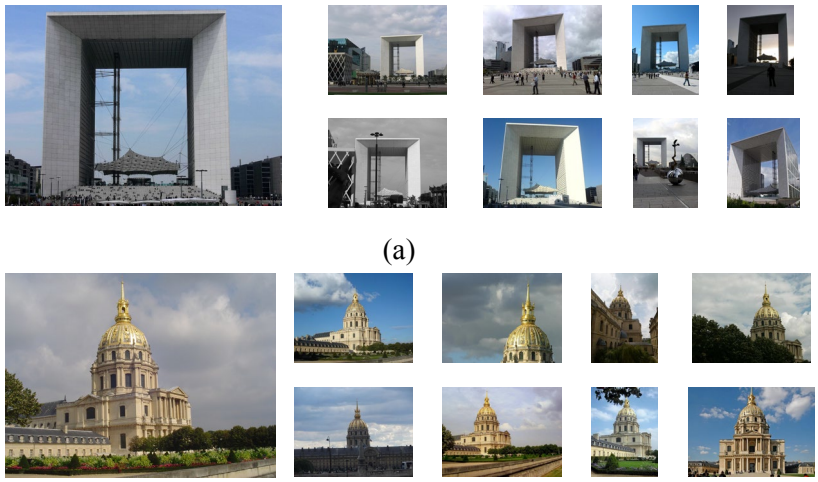
In past decades, many instance retrieval systems [Jégou, Douze and Schmid (2010); Mikulik, Perdoch, Ondřej et al. (2013); Philbin, Chum, Isard et al. (2007); Arandjelovic and Zisserman (2012); Tao, Gavves, Snoek et al. (2014); Lew (2006)] are based on the Bag-of-Word (BoW) model with hand-crafted local features (e.g., SIFT [Lowe (2004)], SURF [Bay, Ess, Tuytelaars et al. (2008)]). The local features are quantized to the nearest visual word of a well-trained visual codebook, and the inverted index is built for efficient feature matching. The BOW-based retrieval framework has shown to be suitable for large-scale instance search tasks.

[1] Jiangsu Engineering Center of Network Monitoring & Department of Computer and Software, Nanjing University of Information Science and Technology, Nanjing, 210044, China.

[2] Department of Electrical and computer Engineer, University of Winsor, N9B3P4, Winsor, ON, Canada.

[3] Department of Computer Science and Information Engineering National Dong Hwa University, Shoufeng, Hualien 974, Taiwan.

[*] Corresponding Author: Zhili Zhou. Email: zhou_zhili@163.com.



(a)

(b)

Figure 1: The toy examples of query images(left) and the database images (right) which contain the same target objects

However, due to the limited discriminative power of hand-crafted local features and BOW quantization error, there are many mismatches between images, which will decrease the retrieval accuracy significantly. To alleviate this issue, some post-processing methods are proposed to explore spatial information among visual words, or to adopt query expansion [Arandjelovic and Zisserman (2012); Mohedano, Salvador, Mcguinness et al. (2016); Razavian, Azizpour, Sullivan et al. (2014); Chum, Philbin, Sivic et al. (2007); Yuan and Sun (2018)] to improve the retrieval accuracy.

Recently, Convolutional Neural Networks (CNN) are well known for their out-standing performance in many computer vision tasks, such as image classification, target detection, and instance segmentation. Recent work Babenko et al. [Babenko and Lempitsky (2015)] has shown that features directly extracted from pre-trained CNNs outperform the hand-crafted features for instance retrieval on several public retrieval benchmarks. The existing instance retrieval frameworks using CNN-based image representations can be roughly divided into three categories. 1) The first one directly extracts a global CNN feature from an entire image [Wan, Wang, Hoi et al. (2014)]; 2) The second one extracts the CNN features at the local regions detected from an image, and then employs some aggregation techniques originally designed for aggregating hand-crafted local features, such as VLAD [Jégou, Douze, Schmid et al. (2010)], Fisher vector [Perronnin, Liu, Jorge et al. (2010)], and Triangulation embedding [Jegou and Zisserman (2014)], to further aggregate these features into a global feature, e.g., Gong et al. [Gong, Wang, Guo et al. (2014); Tolias, Sicre and Jégou (2015); Mohedano, Salvador, Mcguinness et al. (2016); Yuan, Sun and Wu (2018)]. 3) The last one builds a new network model to con-duct end-to-end learning process to learn image representation, such as Liu et al. [Liu, Tian, Wang et al. (2016); Jimenez, Alvarez and Giro-I-Nieto (2017); Liu, Luo, Qiu et al. (2016); Arandjelovic, Gronat, Torii et al. (2017)].

The task of instance retrieval is to search for the images containing a same target object of a given query image in a large-scale database. However, the target objects are usually

contaminated by many irrelevant patterns and background clutter, as shown by the toy examples in Fig. 1. The existence of the irrelevant patterns and background clutter make the task of instance retrieval quite challenging. Unfortunately, most of the existing CNN features for instance retrieval are global features, which describe the patterns of the entire images, leading to inferior performance for retrieving target objects.

In order to effectively identify the target objects from the irrelevant image patterns and background clutter, it is more reasonable to extract CNN features at image region-level and cross-match these region-level CNN features between images. Therefore, we attempt to propose a RoI-based CNN feature for instance retrieval. First, it detects region-of-interests (RoIs) based on the properties of convolutional layer activations of a pre-trained CNN, and then extracts the CNN features of ROIs from the fully-connected layer.

Finally, the RoI-based CNN features are cross-matched between images by an efficient feature matching strategy. We test the performances of the proposed approach, when using different fully-connected or convolutional layers of the famous pre-trained network model, i.e., Alexnet [Krizhevsky, Sutskever and Hinton (2012)], VGG16 [Simonyan and Zisserman (2014)], for RoI-based CNN feature extraction. The experimental results tested on several instance retrieval benchmarks validate the effectiveness of our proposed approach.

The contributions of this paper are given as follows. 1) The extraction of the RoI-based features, which can support region-level visual matching and achieve superior performance than the existing global CNN-based features for instance retrieval. 2) The proposed method has great flexibility for region generation in that potential RoIs of any size and aspect ratio can be obtained anywhere in the image. 3) An efficient region-level feature matching strategy is proposed. It can efficiently cross-match the RoI-based features between images.

The rest of this paper is organized as follows. Section 2 describes the related work. Section 3 presents the RoI-based CNN feature extraction method, and the efficient region-level feature matching strategy. Section 4 provides the experimental results of the proposed approach and makes comparison with the state-of-the-arts. The conclusion is given in Section 5.

2 Related work

In past decades, many hand-crafted local features such as SIFT and SURF have been widely adopted in traditional instance retrieval systems. The invariant local features are extracted to describe the low-level characteristics of image patches, such as color or texture information, and they have shown desirable robustness to a variety of common image modifications and distortions, such as rescaling, rotation, illuminance and contrast change. To avoid the exhaustive feature matching between images, the BOW model and classic inverted index structures are adopted to realize fast feature matching in large-scale image databases. The BoW model is usually followed by some post-processing steps, e.g., spatial verification [Li, Jiang, Zha et al. (2013)] or query expansion [Arandjelovic and Zisserman (2012); Mohedano, Salvador, Mcguinness et al. (2016); Razavian, Azizpour, Sullivan et al. (2014); Chum, Philbin, Sivic et al. (2007)]. In the BOW-based frameworks, a large-sized vocabulary is used to significantly improve retrieval quality and it also need considerable memory. To address this issue, some aggregation methods such as VLAD [Jégou, Douze,

Schmid et al. (2010)], Fisher vector [Perronnin, Liu, Jorge et al. (2010)], and Triangulation embedding [Jegou and Zisserman (2014)] are proposed to aggregate the local features of an image into a compact feature.

Recently, as CNN has achieved great success in the field of many computer vision tasks, it is also possible to use the intermediate layer activations of CNN as image representation for instance retrieval. Some works Gong et al. [Gong, Wang, Guo et al. (2014); Babenko, Slesarev, Chigorin et al. (2014); Wan, Wang, Hoi et al. (2014); Razavian, Azizpour, Sullivan et al. (2014); Yuan, Xia, Jiang et al. (2019)] directly extract global or local features from the activations of fully-connected layers of pre-trained CNN. For example, Wan et al. [Wan, Wang, Hoi et al. (2014)] directly used the output of the fully-connected layer from the pre-trained network as image representations. Then, the features were post-processed by l_2 normalization and similarity learning to improve the retrieval accuracy.

Instead of using full-connected layers, some works Babenko et al. [Babenko and Lempitsky (2015); Jimenez, Alvarez and Giro-I-Nieto (2017); Mohedano, Salvador, Mcguinness et al. (2016); Rezende, Zepeda, Ponce et al. (2017)] prefer to extract image representations from the convolutional layer activations. Generally, they generate the activations of deep convolutional layers with an input image, and then aggregate these activations by performing spatial max-pooling [Tolias, Sicre and Jégou (2015)], sum-pooling [Babenko and Lempitsky (2015); Jimenez, Alvarez and Giro-I-Nieto (2017)], or mean-pooling [Zhi, Duan, Wang et al. (2016)] to generate compact image representations. To further improve the retrieval accuracy, these representations are usually post-processed by l_2 normalization and PCA whitening. In these works, the image representations generated from convolutional layers have been shown superior performance than those from full-connected layers for instance retrieval. However, target objects in an image typically occupy only a small proportion of an image, while the above image representations are extracted to describe the entire image. That makes these representations less robustness to background clutters, leading to inferior performance for identifying target objects from distracters in cluttered backgrounds.

Recently, regional CNN features [Gong, Wang, Guo et al. (2014); Tolias, Sicre and Jégou (2015); Mohedano, Salvador, Mcguinness et al. (2016); Hinami, Matsui and Satoh (2017)] have shown significantly advantages over global CNN features in that they represent an image through a set of regions. Then, the regional CNN features can be cross-matched between images to implement instance retrieval. In such manner, the problem of robustness to background clutter can be alleviated significantly. For example, Razavian et al. [Razavian, Azizpour, Sullivan et al. (2014)] first investigated the use of regional CNN features in instance retrieval. Fischer et al. [Fischer, Dosovitskiy and Brox (2014)] detected the elliptic regions of interest by using the maximally stable extremal regions (MSER) detector. Then, the CNN and SIFT features are extracted from these regions. Experimental results illustrate the significant advantages of CNN over SIFT. Although the above region-based CNN features generally have achieved superior performance than the global CNN features, there still suffer the following shortcoming. Generally, they are generated from the rectangular blocks divided from the entire image or the small patches detected by the traditional patch detectors, and many background clutters are introduced into the regions. As a result, these regional features cannot be accurately matched for instance retrieval.

Some other methods are proposed to employ fine-tuning network structure to accommodate instance retrieval tasks. Liu et al. [Liu, Luo, Qiu et al. (2016)] re-adjusted the network structure so that the network can simultaneously output multiple values to predict multiple attributes of an image instance. Liu et al. [Liu, Tian, Wang et al. (2016)] proposed an end-to-end learning framework to identify identical vehicles in different images by capturing both the inter-model difference and intra-model difference between different vehicles. Jimenez et al. [Jimenez, Alvarez and Giro-I-Nieto (2017)] modified the network structure and replaced the fully-connected layer with the global average pool layer. They represent images by aggregating pooling features corresponding to the top N categories of the highest prediction.

In this paper, we propose a RoI-based CNN feature using convolutional feature maps (CFMs) [Cao, Liu, Wang et al. (2016)] for instance retrieval. Obviously, it is very likely that the visual patterns with similar texture information belongs to the same object. That is also illustrated by Fig. 2.

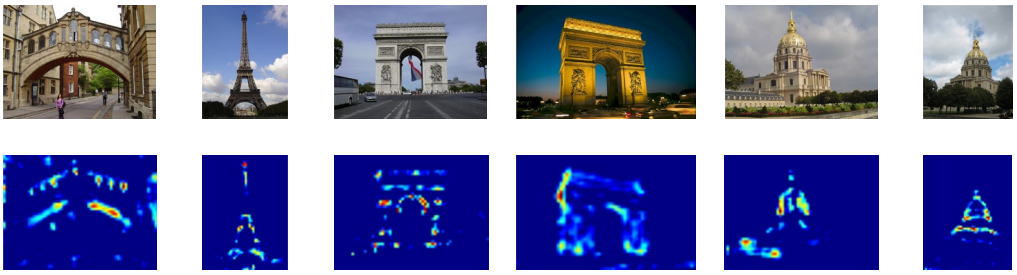


Figure 2: The feature maps produced by a certain layer (conv5) of Alexnet with input images. The first row shows the input images, and the second row visualizes the corresponding feature maps

This figure visualizes the activation values of a certain layer of CNN with an input image, and these activations are viewed as the responses of a certain convolutional filter. From this figure, we observe that the visual patterns with the similar activation values belong to the same object. Based on this observation, we detect multiple RoIs, i.e., potential target object regions, from an image based on its CFMs properties. Then, the regional CNN features are extracted from the CFMs of these regions. Finally, the instance retrieval is implemented by comparing these RoI-based CNN features between images.

3 The proposed ROI-based CNN features extraction method

In this section, we detail the extraction process of RoI-based CNN features and introduce how to use these features for instance retrieval.

3.1 Generation of convolutional feature maps (CFMs)

In our approach, the famous pre-trained network, i.e., Alexnet, is employed. From Krizhevsky et al. [Krizhevsky, Sutskever and Hinton (2012)], it has five convolutional layers (conv1, conv2, ..., conv5) followed by three fully connected layers (fc6, fc7, fc8). If an image is fed into a pre-trained CNN, the output of a convolutional layer is a set of CFMs, which is a 3D tensor of size W×H×D. Where D is the number of output feature channels, and W and H are

proportional to the width and height of the input image, respectively.

3.2 RoI-based feature extraction

In this subsection, we introduce how to generate base regions from images. The base region generation consists of two steps: base region detection based on CFMs and base region optimization.

3.2.1 Base region detection based on CFMs

Intuitively, the CFMs of CNN pre-trained on large labeled datasets have different responses for different textures. Generally, in the CFMs, high activation values correspond to well-textured region, while low activation values represent weakly textured regions. Based on this phenomenon, we can detect relatively high activation values, and combine the positions of these activation values to form a potential target region in an image.

For each map in CFMs, the binarization operation is performed by a defined threshold to obtain the corresponding binarized map. Then, for $d-th$ map in the binarized CFMs, the locations of non-zero values are utilized to form a base region, which is represented by

$$R_d = \{P_i^d | f(P_i^d) > 0\}, 1 \leq d \leq D \quad (1)$$

where, $f(P_i^d)$ is the activation value of P_i^d in $d-th$ feature map, and D is the number of feature maps. Afterward, for simplification, the rectangle regions R'_d are used as base regions to approximate the detected irregular regions, as shown in Fig. 3.

As a result, totally D base regions are generated from an image.

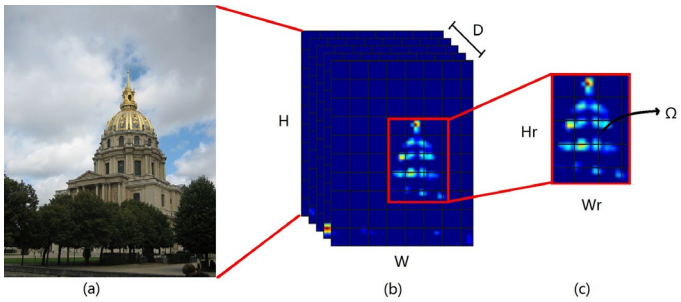


Figure 3: The toy examples of (a) query image and (b) its CFMs, which are a 3D tensor of size W×H×D. (c) A base region extracted from locations of non-zero activations in a feature map

3.2.2 RoI detection

In the generated base regions, there are two kinds of regions that are less likely to be the regions of target objects. In this subsection, we will eliminate these regions to optimize the result of base region generation. The first kind of regions are those regions with relatively small size. If the ratio between the areas of the regions and that of the entire image is less than the threshold T_{min}, we will remove these regions. The second kind is the regions of

which aspect ratio ($= \frac{\max(W_r,H_r)}{\min(W_r,H_r)}$) is greater than the predefined threshold T_{rate}, where W_r and H_r represent the width and length of the base region, respectively. Next, the sum of the activation values for each remaining base region is calculated and sorted in descending order. Then, we iteratively filter the base regions until we select K regions with higher activations, meanwhile, we guarantee that the Intersection over Union (IOU) between regions are less than T_{IOU} (T_{IOU} is set to 95%).

Finally, the RoIs of an image are obtained according to the relationships between the sizes of the image and its CFMs. Fig. 4 visualizes the ROIs generated from some example images using CFMs. Fig. 5 shows the extraction pipeline of RoIs from a given image.

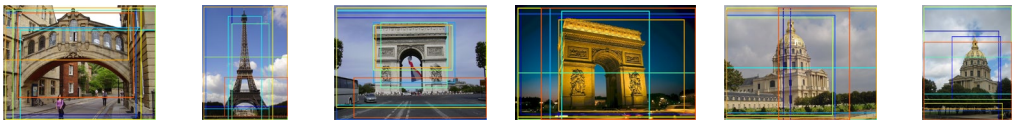


Figure 4: The base regions generated from different instance images. These regions are shown by the bounding boxes with different color

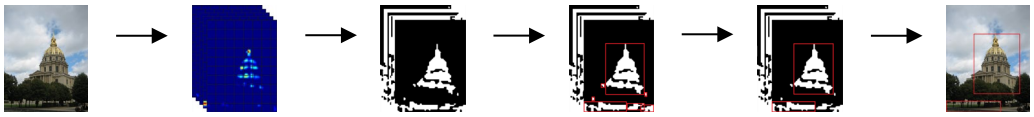


Figure 5: The extraction of RoI-based features from a given image. (a) The input image; (b) The Corresponding CFMs; (c) The binarized CFMs; (d) The base regions generated from the binarized CFMs; (e) The remaining base regions after region optimization; (f) The RoIs of the input image generated by mapping the base regions to the image

3.2.3 Feature extraction

We extract two kinds of CNN features to represent ROIs, i.e., one is from the fully-connected layer and the other from the convolutional layer.

The feature generated from the fully-connected layer usually have a higher level of semantic abstraction for the objects in the image, which allows high discriminability for distinguishing different objects. For a ROI in the image, we feed it directly into Alexnet to get the output of the fully-connected layer as the feature of this region. Then, PCA-whitening and l_2 normalization are implemented to obtain compact feature representation.

Recent work has shown that the features of convolutional layers have better discriminating power than fully-connected layers. As the feature extracted from the convolutional layer preserve spatial information of the objects in the image, we also extract the convolutional layer feature. The convolutional layer feature of the ROI is extracted by sum-pooling or max-pooling of the corresponding portion of CFMs. Consequently, the feature dimensionality is equal to the number of feature channels.

The binarization is performed to all the extracted CNN features. We perform binarization to a feature f_i by

$$f_i = \begin{cases} 1, & if\ f_i > mean(f_i) \\ 0, & \text{otherwise} \end{cases} \quad (2)$$

3.2.4 Image retrieval and re-ranking

We use a combination of low-level visual features and semantic features to represent each image. For each image in a database, the SURF descriptors are extracted as the low-level features and quantized to visual words based on BOW model, and thus the image can be represented as a histogram of visual words, i.e., BOW representation. Then, the classic inverted file is built for fast retrieval. At the same time, the CNN features of each RoI in the image are retained as high-level semantic features.

In the query stage, we first match the quantized SURF features between a given query image and database images by looking up the inverted index file, and then score the similarities between database images to the query image by TF-IDF strategy. Then the top N ranked database images are used as candidate images for re-ranking of initial retrieval result. During the query re-ranking stage, RoI-based features are extracted from the query image by the same feature extraction algorithm described above.

Then, the RoI-based CNN features are cross-matched between query and database images, and the maximum feature similarities are used as the similarity between them. To improve the matching efficiency, we propose a fast feature matching method to avoid exhaustive matching between images. In the matched CNN features between query image and top n candidate images where n≪N, the true feature matches usually occupy a large proportion. Based on this observation, in the query re-ranking stage, we use the exhaustive feature matching only for the top n candidate images, and thus obtain the first k RoI-based features with highest matching frequencies in the query image. Thus, the query image can be accurately represented by these top k RoI-based CNN features. Afterword, for the remaining (N-n) candidate images, we only calculate the similarity between their RoI-based CNN features and the first k RoI-based features of the query image. This simple fast matching strategy can significantly reduce the computational complexity of feature matching, and also suppress the similarity between dissimilar images, which will improve retrieval accuracy.

Finally, the maximum feature similarity between query image and candidate images is used as the image similarity for the re-ranking of initial retrieval result.

4 Experiments

In this section, we will systematically evaluate the performance of the proposed method on public image datasets, and compare with those of the state-of-the-art methods.

4.1 Datasets

We evaluate our approach on the following datasets.

Oxford5k dataset. This dataset contains 5062 images of 11 landmarks from Oxford University. Each landmark contains 5 query images, and thus there are 55 queries in total.

Paris6k dataset. This dataset is similar to Oxford5k, but it contains 6412 images of 11 Paris

land-marks. Each landmark contains 5 query images, and 55 queries in total.

To reduce computation complexity, the resolution of all database images is adjusted to no more than 500×500 pixels.

4.2 Experiment setup

Mean average precision (mAP) is adopted to compute the accuracy of the proposed method.

We find that the features extracted from conv5 layer of CFMs are more effective than other convolutional layers, and the features extracted from the fc7 layer provide superior performance to those of other layers. Therefore, in the remaining experiments, we extract RoIs and then use the features extracted from fc7 or conv5 layer to represent each RoI. All experiments are performed on a PC with 3.2 GHz Core-i5 (8 GB RAM).

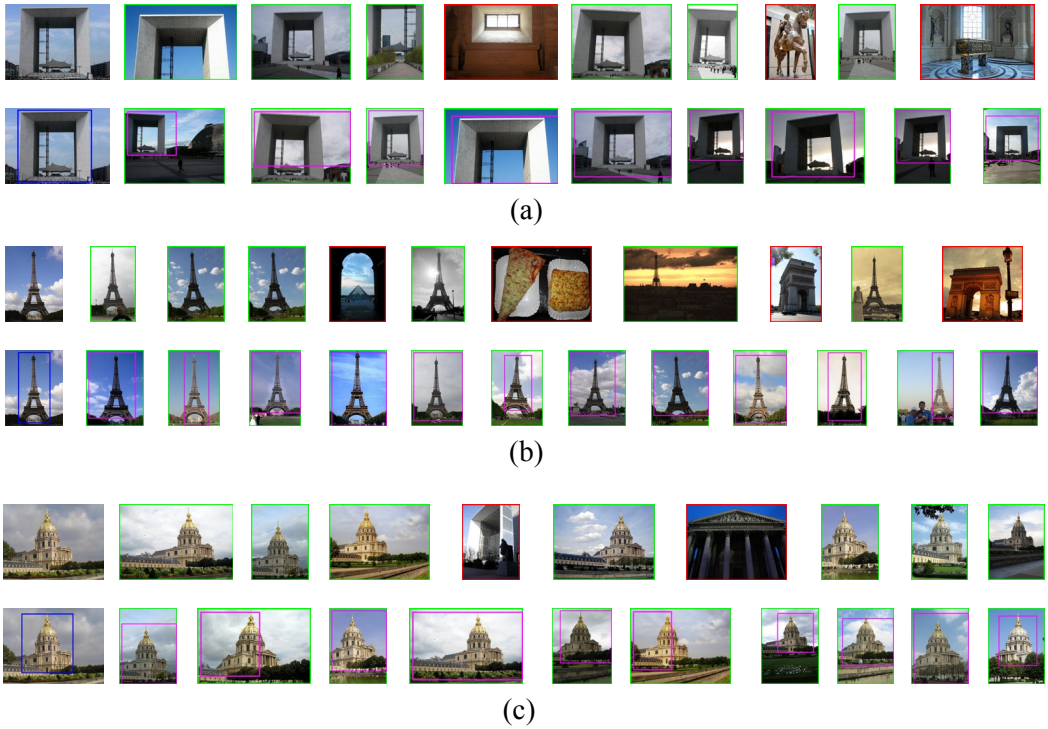


Figure 6: Examples of top retrieved images before (top) and after (bottom) re-ranking with RoI-based feature extraction. The left is query images in which the objects are highlighted by blue box. The best matching object position after re-ranking is highlighted by bounding box with magenta color. The positive/negative images are marked with green/red border

Fig. 6 is a query result for some query images. The left column is the query image, the first row on the right is the top retrieved images returned by using the BoW model, and the second row is the top retrieved results returned by the re-ranking. The true positives are marked by the green color, and the false positives are shown by red color.

For vector f_1 and f_2 $(f_1, f_2 \in R^{1*d})$, the similarity is computed as

$$\mathrm{Sim}(f_1, f_2) = 1 - \frac{sum(abs(f_1 - f_2))}{D} \quad (3)$$

where D represents the feature dimensionality.

We only analyze the impact of different K with small values (K=5,10,15,20) on the retrieval results, as higher K leads to an exponential increase in the computation cost of similarity. Tab. 1 and Tab. 2 show the retrieval accuracy of the proposed method using different K and feature extraction methods. As shown in Tab. 1 and Tab. 2, larger K leads to higher retrieval accuracy. We can also see that the features extracted from the fully-connected layer can yield better results than those from the convolutional layer. The reason may be that the features from the fully-connected layer contain a higher level of visual abstraction, while features from the convolutional layer tend to capture low-level information of the target object. In the two tables, Dim represents the dimensionality of the feature. Baseline refers to the retrieval accuracy using only global image representation; Fc7 corresponds to the activation value of the fc7 layer directly from the CNN; Max-pooling or Sum-pooling means that max-pooling or sum-pooling is performed for the conv5 feature maps.

Table 1: Retrieval result on the Paris6k dataset

Feature	Dim	Baseline	K=5	K=10	K=15	K=20
Fc7	4096	80.55	85.04	85.93	**86.78**	**86.78**
Max-pooling	256	81.64	82.4	81.63	81.84	81.01
Sum-pooling	256	77.15	82.38	81.55	81.24	80.51

Table 2: Retrieval result on the Oxford5k dataset

Feature	Dim	Baseline	K=5	K=10	K=15	K=20
Fc7	4096	65.42	71.23	72.32	**72.76**	**72.86**
Max-pooling	256	69.21	71.09	71.33	69.62	68.34
Sum-pooling	256	65.31	71.76	71.99	71.31	69.32

4.3 Comparison with the state-of-the-arts

In this section, we compare RoI-based feature extraction with the state-of-the-art methods on several instance retrieval datasets. From Tab. 3, it clear that our method achieves highest MAP values on Paris6K dataset. Also, the MAP value of our method is comparable to that of Mohedano et al. 's method [Mohedano, Salvador and Mcguinness (2016)].

Table 3: Comparison with existing methods based on CNN-based extraction

Methods	NetWork	Layer	Paris6k	Oxford5k
[Mohedano, Salvador, Mcguinness et al. (2016)]	VGG16	conv	84.8	**78.8**
[Razavian, Azizpour, Sullivan et al. (2014)]	OverFeat	fc	67.6	52.0
[Ng, Yang and Davis (2015)]	VGG16	conv	69.4	64.9
[Mopuri and Babu (2015)]	Alexnet	fc	71.47	60.71
[Babenko and Lempitsky (2015)]	VGG16	conv	-	65.7
Fc7 (ours)	Alexnet	fc	**86.78**	72.86

5 Conclusion

In this paper, we propose a novel RoI-based feature extraction method. By analyzing the properties of CFMs of images, multiple ROIs that may contain the target object are detected. Using these regions for the extraction of image feature representations can effectively reduce the impact of image background on retrieval performance. Moreover, the proposed method can be easily combined with other post-processing algorithms to further improve the retrieval performance. Experiments demonstrate the effectiveness of the proposed method in re-ranking of initial retrieval result on several public datasets.

Acknowledgments: This work is supported by the National Natural Science Foundation of China under Grant 61602253, U1836208, U1536206, U1836110, 61672294, in part by the National Key R&D Program of China under Grant 2018YFB1003205, in part by the Priority Academic Program Development of Jiangsu Higher Education Institutions (PAPD) fund, in part by the Collaborative Innovation Center of Atmospheric Environment and Equipment Technology (CICAEET) fund, China, and in part by MOST under contracts 108-2634-F-259-001- through Pervasive Artificial Intelligence Research (PAIR) Labs, Taiwan.

References

Arandjelovic, R.; Gronat, P.; Torii, A.; Pajdla, T.; Sivic, J. (2017): NetVLAD: CNN architecture for weakly supervised place recognition. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, pp. 5297-5307.

Arandjelovic, R.; Zisserman, A. (2012): Three things everyone should know to improve object retrieval. *IEEE Conference on Computer*.

Babenko, A.; Lempitsky, V. (2015): Aggregating deep convolutional features for image retrieval. *Computer Science*, *Vision and Pattern Recognition, Computer Society*, pp. 2911-2918.

Babenko, A.; Slesarev, A.; Chigorin, A.; Lempitsky, V. (2014): Neural codes for image retrieval. *European Conference on Computer Vision*, vol. 8689, pp. 584-599.

Bay, H.; Ess, A.; Tuytelaars, T.; Gool, L. V. (2008): Speeded-up robust features (SURF). *Computer Vision & Image Understanding*, vol. 110, no. 3, pp. 346-359.

Cao, J.; Liu, L.; Wang, P.; Huang, Z.; Shen, C. et al. (2016): Where to focus: query adaptive matching for instance retrieval using convolutional feature maps. arXiv:1606.06811.

Chum, O.; Philbin, J.; Sivic, J.; Isard, M.; Zisserman, A. (2007): Total recall: automatic query expansion with a generative feature model for object retrieval. *11th International Conference on Computer Vision*, pp. 1-8.

Fischer, P.; Dosovitskiy, A.; Brox, T. (2014): Descriptor matching with convolutional neural networks: a comparison to SIFT. *Computer Science*.

Gong, Y.; Wang, L.; Guo, R.; Lazebnik, S. (2014): Multi-scale orderless pooling of deep convolutional activation features. *Computer Vision and Pattern Recognition*, vol. 8695, pp. 392-407.

Hinami, R.; Matsui, Y.; Satoh, S. (2017): Region-based image retrieval revisited. http://www.doc88.com/p-6199188690897.html.

Jégou, H.; Douze, M.; Schmid, C. (2010): Improving bag-of-features for large scale image search. *International Journal of Computer Vision*, vol. 87, no. 3, pp. 316-336.

Jégou, H.; Douze, M.; Schmid, C.; Perez, P. (2010): Aggregating local descriptors into a compact image representation. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 238, no. 6, pp. 3304-3311.

Jegou, H.; Zisserman, A. (2014): Triangulation embedding and democratic aggregation for image search. *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3310-3317.

Jimenez, A.; Alvarez, J. M.; Giro-I-Nieto, X. (2017): Class-weighted convolutional features for visual instance search. arXiv:1707.02581.

Krizhevsky, A.; Sutskever, I.; Hinton, G. (2012): ImageNet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, vol. 25, no. 2, pp. 1097-1105.

Lew, M. S. (2006): Content-based multimedia information retrieval: State of the art and challenges. *ACM Transactions on Multimedia Computing Communications and Applications*, vol. 2, no. 1, pp. 1-19.

Li, L.; Jiang, S.; Zha, Z. J.; Wu, Z.; Huang, Q. (2013): Partial-duplicate image retrieval via saliency-guided visual matching. *IEEE Multimedia*, vol. 20, no. 3, pp. 13-23.

Liu, H.; Tian, Y.; Wang, Y.; Pang, L.; Huang, T. (2016): Deep relative distance learning: tell the difference between similar vehicles. *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2167-2175.

Liu, Z.; Luo, P.; Qiu, S.; Wang, X.; Tang, X. (2016): DeepFashion: powering robust clothes recognition and retrieval with rich annotations. *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1096-1104.

Lowe, D. G. (2004): Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91-110.

Mikulik, A.; Perdoch, M.; Ondřej, C.; Matas, J. (2013): Learning vocabularies over a fine quantization. *International Journal of Computer Vision*, vol. 103, no. 1, pp. 163-175.

Mohedano, E.; Salvador, A.; Mcguinness, K.; Marques, F.; O'Connor, N. E. et al. (2016): Bags of local convolutional features for scalable instance search. *ACM on International Conference on Multimedia Retrieval*, pp. 327-331.

Mopuri, K. R.; Babu, R. V. (2015): Object level deep feature pooling for compact image representation. *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 62-70.

Ng, Y. H.; Yang, F.; Davis, L. S. (2015): Exploiting local features from deep networks for image retrieval. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 53-61.

Perronnin, F.; Liu, Y.; Jorge, S.; Poirier, H. (2010): Large-scale image retrieval with compressed Fisher vectors. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 3384-3391.

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

Razavian, A. S.; Azizpour, H.; Sullivan, J.; Carlsson, S. (2014): CNN Features off-the-shelf: an astounding baseline for recognition. *Computer Vision and Pattern Recognition*, pp. 512-519.

Rezende, R.; Zepeda, J.; Ponce, J.; Bach, F.; Perez, P. (2017): Kernel square-loss exemplar machines for image retrieval. *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7263-7271.

Simonyan, K.; Zisserman, A. (2014): Very deep convolutional networks for large-scale image recognition. *Computer Science*.

Tao, R.; Gavves, E.; Snoek, C. G. M.; Smeulders, A. W. M. (2014): Locality in generic instance search from one example. *Computer Vision and Pattern Recognition*, pp. 2099-2106.

Tolias, G.; Sicre, R.; Jégou, H. (2015): Particular object retrieval with integral max-pooling of CNN activations. *Computer Science.*

Wan, J.; Wang, D.; Hoi, C. H.; Wu, P.; Zhu, J. et al. (2014): Deep learning for content-based image retrieval: a comprehensive study. *International Conference on Multimedia*, pp. 157-166.

Yuan, C.; Sun, X. (2018): Fingerprint liveness detection using histogram of oriented gradient based texture feature. *Journal of Internet Technology*, vol. 19, no. 5, pp. 1499-1507.

Yuan, C.; Sun, X.; Wu, Q. M. J. (2018): Difference co-occurrence matrix using BP neural network for fingerprint liveness detection. *Soft Computing*, pp. 1-13.

Yuan, C.; Xia, Z.; Jiang, L.; Cao, Y.; Wu, Q. M. et al. (2019): Fingerprint liveness detection using an improved CNN with image scale equalization. *IEEE Access*.

Zhi, T.; Duan, L. Y.; Wang, Y.; Huang, T. (2016): Two-stage pooling of deep convolutional features for image retrieval. *International Conference on Image Processing*, pp. 2465-2469.