Image Retrieval Research Based on Significant Regions

Jie Xu(✉), Shuwei Sheng, Yuhao Cai, Yin Bian, and Du Xu

School of Information and Communication Engineering, University of Electronic Science and Technology of China, Chengdu, China
xuj@uestc.edu.cn

Abstract. Deep Convolution neural networks (CNN) has achieved great success in the field of image recognition. But in the image retrieval task, the global CNN features ignore local detail description for paying too much attention to semantic information of images. So the MAP of image retrieval remains to be improved. Aiming at this problem, this paper proposes a local CNN feature extraction algorithm based on image understanding, which includes three steps: significant regions extraction, significant regions description and pool coding. This method overcomes the semantic gap problem in traditional local characteristic and improves the retrieval effect of global CNN features. Then, we apply this local CNN feature in the image retrieval task, including the same category retrieval task by feature fusion strategy and the instance retrieval task by re-ranking strategy. The experimental results show that this method has achieved good performance on the Caltech 101 and Caltech 256 classification datasets, and competitive results on the Oxford 5k and Paris 6k instance retrieval datasets.

Keywords: Significant regions · Image understanding · CNN
Image retrieval

1 Introduction

Content based image retrieval (CBIR) uses the description of image content to search similar images. Most of the existing methods employ low-level visual features of image, such as Sift [1], BoW [2], Fisher vector [3] and VLAD [4]. Although CBIR in the past decade has made a lot of scientific research and set up some research or commercial image retrieval systems, but most of the image retrieval performance cannot satisfy the requirement. The main reason is semantic gap problem.

Image descriptors based on the activations within deep convolutional neural networks have emerged as state-of-the-art generic descriptors for visual recognition [5–7]. CNN global characteristics as a kind of high-level semantic representation, used in other recognition tasks and performed well [8–11]. Razavian [12] studied the characteristics of global CNN feature and used it for different image recognition tasks, including image retrieval. Yandex [13] achieve a remarkable increase in performance by fine-tuning CNN model on target dataset and extracting fc6 layer features. Lin [14] utilized hash code to transform fc6 layer features into a binary sequence, and greatly improved the retrieval efficiency.

X. Liu et al. (Eds.): ChinaCom 2018, LNICST 262, pp. 127–136, 2019.
https://doi.org/10.1007/978-3-030-06161-6_12

But there are still some problems to be solved. Global CNN feature contains too many high-level semantic information related to the classification, so it tends to ignore the details of images. Recently, some studies began to focus on the image characteristics of granular to improve the global CNN feature. Wang [15] put forward using triplet to increase similarity intra-class and distinction between classes, and proposed a multi-scale network to increase the local detail information in the image. An adaptive region detection method [16] is proposed to eliminate the street snap clothing pictures and store clothing pictures. Clothing attribute dataset is used to mining the fine-grained properties. CKN [17] network was proposed to extract local degeneration characteristics of images. Then Mattis [18] used unsupervised training CKN network to extract local features in image retrieval task.

Combined the advantage of traditional local features, we propose a local CNN features extraction method based on image understanding. This work presents three contributions:

First, we propose a local CNN feature method including three steps: significant regions extraction, significant regions description and coding. Among them, the significant regions are extracted based on image understanding, which can describe the whole attribute of the image and the relation between different entities.

Figure 1 shows eight extraction results of the significant regions. The raw pictures come from four different datasets, which will be used in the retrieval tasks in Sect. 3.

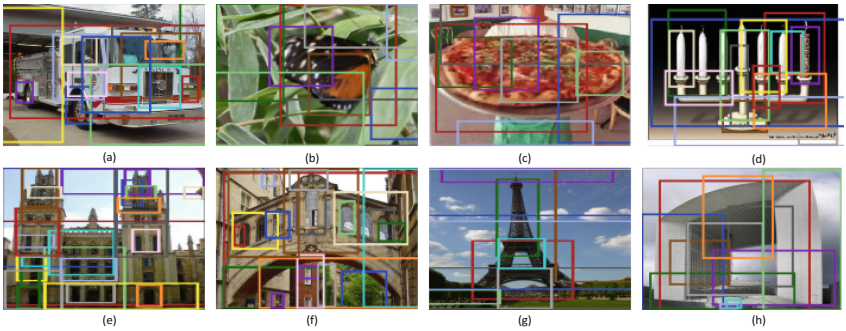


Fig. 1. The results of significant regions. (a) and (b) come from Caltech 101 dataset. (c) and (d) come from Caltech 256 dataset. (e) and (f) come from Oxford 5k dataset. (g) and (h) come from Paris building 6k dataset

Second, we fusion the global CNN feature and local CNN feature and apply it in the same category retrieval task.

Third, we put forward the re-ranking algorithm based on significant regions and employ it to instance retrieval task. The experimental results show that the proposed methods can further improve the accuracy of image retrieval.

The rest of this paper is organized as follows. The details of the proposed local CNN features and image retrieval method are given in Sect. 2. We give the experimental results and analysis in Sect. 3. Finally, we conclude in Sect. 4.

2 Methodology

In this section, we give the details of the proposed local CNN features for image representation and apply it for the image retrieval tasks

2.1 CNN Feature Based on Significant Regions

In traditional image retrieval task, local feature showed greater advantage than global feature, because it can describe more details information and have scale, rotation and brightness invariant. Sift feature is a very common local descriptor. it contains key points detection, key points description and coding three steps to condense the image information into 128 dimensional feature vector. In view of the outstanding characteristics of the sift feature, this paper uses image understanding theories and models to extract the significant regions. Then by significant regions description and coding we generate local CNN feature. Algorithm process as shown in Fig. 2.

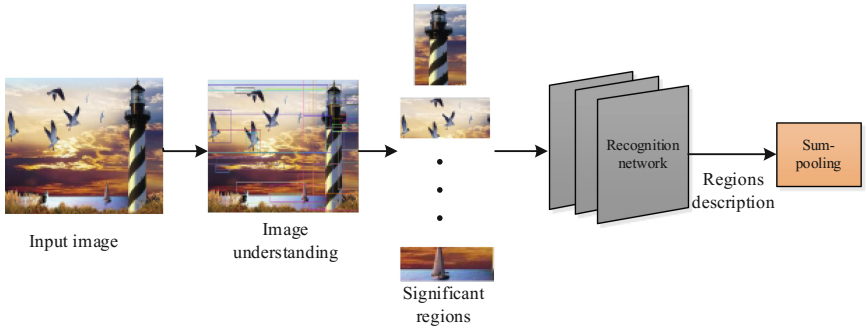


Fig. 2. Local CNN feature extraction process

In Fig. 2, the process of extracting local CNN feature can be divided into four parts

(a) CNN + RPN + LSTM model used to extract the significant regions and the model is trained on image understanding task;
(b) Picking out the highest score of K a significant area;
(c) Describing the K significant regions through identification network and generating feature description;
(d) Coding by sum-pooling.

2.2 Significant Regions Extraction Based on Image Understanding

In order to solve semantic gap problem, this paper attempts extract significant area from the perspective of image understanding

In image caption task, it needs to locate the target area, and also describes the target area in natural language. We use CNN + RPN + LSTM structure to locate the

significant regions, then filter these areas and code to generate low dimensional feature vector. The model structure is shown in Fig. 3.

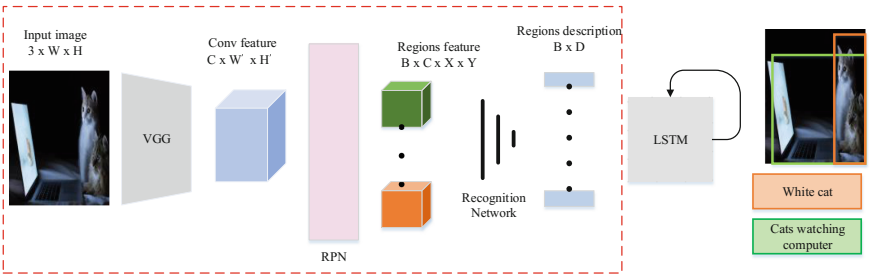


Fig. 3. CNN + RPN + LSTM model

CNN network. We discard the softmax layer and fully connected layers of the original network. Given an input image I of size W × H, the activations (responses) of a convolutional layer form a 3D tensor of W' ×H' × C dimensions, where C is the number of output feature channels, In this paper, C = 512, W' = ⌊W/16⌋, H' = ⌊H/16⌋.

RPN localization layer. RPN localization layer receives the input feature maps, then pinpoints the interest regions and extracts an appropriate length denoting from every region. The structure of RPN localization layer based on the idea of Faster R-CNN. We replace the ROI mechanism in Faster R-CNN to bilateral interpolation method, which can makes candidate regions back propagate the edge information to previous layers. So the edge information can be learned in the process of training. The localization layer accepts a tensor of activations of size C × W' × H'. It then internally selects B regions of interest and returns three output tensors giving information about these regions:

- Region Coordinates: A matrix of shape B × 4 giving bounding box coordinates for each output region.
- Region Scores: A vector of length B giving a confidence score for each output region. Regions with high confidence scores are more likely to correspond to ground-truth regions of interest.
- Region Features: A tensor of shape B × C × X × Y giving features for output regions.

RPN layer is mainly to locate the candidate regions, and filter the regions according to NMS. The rest of the regions are the significant regions in this paper

Recognition network. The recognition network is a fully-connected neural network that processes region features from the localization layer. The features from each region are flattened into a vector and passed through two full-connected layers, each using rectified linear units and regularized using dropout. For each region this produces a description of dimension D = 4096 that compactly encodes its visual appearance. This description of dimension B × 4096 is what we need to code in our task.

LSTM language model. The LSTM model only work in the model training process. We just use it to make the model oriented to image understanding task rather than classification task. We use the Visual Genome dataset [19] to pre-train the model. The aim is to make the model have the ability to locate the significant regions and dig the relation between different entities. The description we need is from recognition network.

2.3 Sum Pooling Coding

In the previous steps, we complete preliminary coding through the recognition network. Then the description of dimension B × 4096 will need to encode into a feature vector in image retrieval task. And the principle for the sum pooling coding is:

First of all, we calculate the sum of feature value and in all significant regions about each dimension

$$F_k^{'} = \sum_{i=1}^{B} C_i \# \tag{1}$$

Then the feature code is the proportion of each dimension

$$F_k = \frac{F_k^{'}}{\sum_{k=1}^{4096} F_k^{'}} \# \tag{2}$$

2.4 Same Category Retrieval

This article takes the algorithm of fusion global CNN features and local CNN in the same category retrieval task. In this part, our local CNN feature based on significant regions aims to improve the global CNN global feature which cannot describe the local details of the image. Algorithm process is shown in Fig. 4.

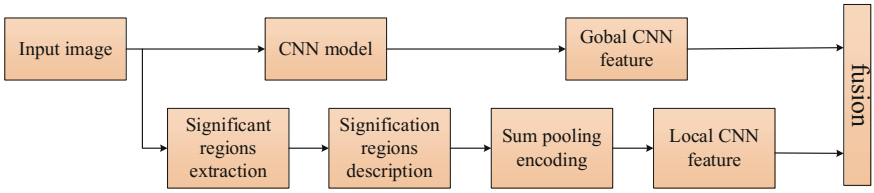


Fig. 4. The process of same category retrieval

In Fig. 4, the upper part employs CNN model as global feature extractor. The second half part is the process of local CNN feature in this paper. Finally, we fuse them together in the same category retrieval task. The dimension of global and local CNN feature vector is 4096.

We use PCA and L2 regularization to fuse them and get the final feature vector. According to the above method, all the pictures can be extracted feature vectors and build features library.

2.5 Instance Retrieval

In instance retrieval, we concentrate on objects in images rather than the class of full image. We employ feature aggregation by cross-dimension weighting proposed by Crow [20] to make initial retrieval and obtain Top-N retrieval results. We propose salient region extraction algorithm to re-rank the Top-N images. The re-ranking algorithm is as follows:

(a) We employ CNN network to extract global query feature q.
(b) We extract salient region in Top-N image and global feature p in every salient region. We re-rank the Top-N based on the similarity between q and p. Images with the highest scores ranking move forward, to further improve the retrieval result.

3 Experiments and Results

3.1 Datasets

We use Caltech 101 and Caltech 256 datasets to verify the results of our method for the same category retrieval task.

Then, we use Oxford Buildings and Paris Buildings to verify the results of our method for the instance retrieval task.

3.2 Experiment in Same Category Retrieval

In this part, we evaluate the ranking of Top-K images with respect to query image q by a precision

$$\mathrm{AP}@k = \frac{\sum_{i=1}^{k} Rel(i)}{k} \tag{3}$$

Where Rel(i) denotes the ground truth relevance between a query q and the i-th ranked image. Here, we consider only the category label in measuring the relevance so Rel(i) $\in$ {0,1} with 1 for the query and the i-th image with the same label and 0 otherwise.

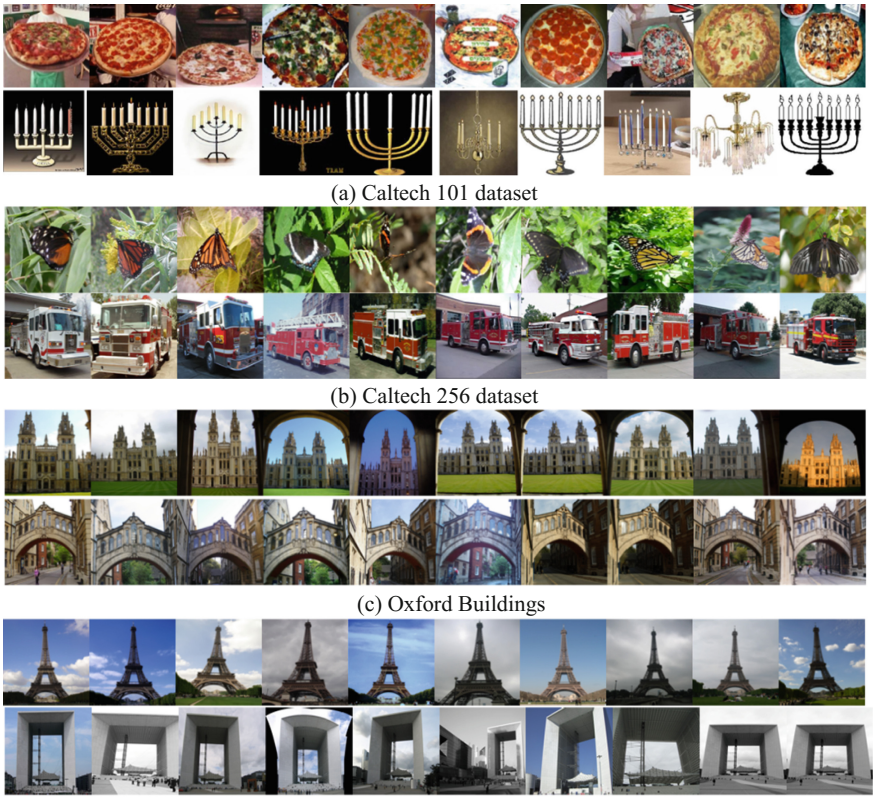
For each dataset in the experiment, the 5% of the total images are randomly selected as query images. When Top-10 retrieval results are returned, the experimental results are shown in Table 1.

Seen from Table 1, the MAP of traditional algorithms such as BoW is low in the same category retrieval task. The image retrieval algorithm based on deep learning achieves better performance. On Caltech 101, the average retrieval accuracy of global CNN features is 79.14%. The average retrieval accuracy of this algorithm is 80.67%, and the average retrieval accuracy of this algorithm is improved by 1.53%. On the

Table 1. The MAP on Caltech 101 and Caltech 256 datasets using different features

Method	Caltech 101	Caltech 256
BoW	0.223	0.268
Global CNN feature	0.791	0.649
Ours	0.807	0.674
Ours + QE	0.801	0.701

Caltech 256 dataset, the average retrieval accuracy of global CNN features is 64.91%. The average retrieval accuracy of this algorithm is 67.37% and the retrieval accuracy is 2.46%, which proves the accuracy and effectiveness of this algorithm. Finally, we visualize the retrieval results, and choose top 12 most similar results with the query image for display. The result is shown in Fig. 5.



(a) Caltech 101 dataset

(b) Caltech 256 dataset

(c) Oxford Buildings

(d) Paris Buildings

Fig. 5. A same category retrieval example on Caltech 101 and Caltech 256 datasets. An instance retrieval example on Oxford Buildings and Paris Buildings. The left-most image in each row corresponds to the query.

3.3 Experiment in Instance Retrieval

In accordance with of the algorithm Sect. 2.3, this article uses the Oxford and Paris dataset to evaluate the MAP of instance retrieval task. And we compare with some state-of-art algorithms such as R-MAC [21], SPOC [4], the Crow [20]. The experimental results as shown in Table 2. From the table, we can find that the MAP of the algorithm in this paper still be proved competitive compare to these state-of-art algorithms. Finally, we visualize the retrieval results, and choose top 12 most similar results with the query image for display. The result is shown in Fig. 5.

Table 2. The MAP on Oxford 5k and Paris 6k datasets using different features

Method	Oxford	Paris
Tr. Embeding [22]	0.560	—
Neural Codes [13]	0.435	—
Razavian et al. [23]	0.533	0.670
Sum pooling [22]	0.589	—
R-MAC [21]	0.669	0.830
SPoC [4]	0.561	0.729
Crow [20]	0.657	0.7347
Ours	0.682	0.737
Ours + QE	0.703	0.751

4 Conclusion

This paper presents a local CNN feature algorithm based on significant regions. The method uses Visual Genome dataset to train the model, which aims to extract the local information by image understanding. This method overcomes the semantic gap problem in traditional local characteristic and improves the retrieval effect of global CNN features. The experimental results show that this method has achieved good performance both in the same category retrieval task and instance retrieval task.

Acknowledgements. This work was supported by National Key Research and Development Program (Grant No. 2016YFB0800105), Sichuan Province Scientific and Technological Support Project (Grant Nos. 2016GZ0093, 2018GZ0255), the Fundamental Research Funds for the Central Universities (Grant No. ZYGX2015J009).

References

1. Lowe, D.G.: Object recognition from local scale-invariant features. In: IEEE International Conference on Computer Vision (2001)
2. Sivic, J.: A text retrieval approach to object matching in videos. In: Proceedings of IEEE International Conference on Computer Vision (2003)

3. Perronnin, F., Liu, Y., Sánchez, J., Poirier, H.: Large-scale image retrieval with compressed Fisher vectors. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Workshops, pp. 3384–3391 (2010)
4. Jegou, H., Perronnin, F., Douze, M., Sánchez, J., Perez, P.: Aggregating local image descriptors into compact codes. IEEE Trans. Pattern Anal. Mach. Intell. **34**(9), 1704–1716 (2012)
5. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems (NIPS 2012), pp. 1097–1105 (2012)
6. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition (2014). arXiv:1409.1556
7. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S.: Going deeper with convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Workshops, pp. 1–9 (2015)
8. Zeiler, M.D., Fergus, R.: Visualizing and understanding convolutional networks. In: Proceedings of the European Conference on Computer Vision, pp. 818–833 (2014)
9. Oquab, M., Bottou, L., Laptev, I., Sivic, J.: Learning and transferring mid-level image representations using convolutional neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Workshops, pp. 1717–1724 (2014)
10. Hoang, T., Do, T.T., Tan, D.K.L., Cheung, N.M.: Selective deep convolutional features for image retrieval. In: ACM, pp. 1600–1608 (2017)
11. Xu, J., Shi, C.Z., Qi, C.Z., Wang, C.H., Xiao, B.H.: Unsupervised part-based weighting aggregation of deep convolutional features for image retrieval. In: AAAI2018 (2018)
12. Razavian, A.S., Azizpour, H., Sullivan, J., Carlsson, S.: CNN features off-the-shelf: an astounding baseline for recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Workshops, pp. 512–519 (2014)
13. Babenko, A., Slesarev, A., Chigorin, A., Lempitsky, V.: Neural codes for image retrieval. In: Proceedings of the European Conference on Computer Vision, pp. 584–599 (2014)
14. Lin, K., Yang, H.F., Hsiao, J.H., Chen, C.S.: Deep learning of binary hash codes for fast image retrieval. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Workshops, pp. 25–37 (2015)
15. Wang, J., Song, Y., Leung, T., Rosenberg, C., Wang, J.: Learning fine-grained image similarity with deep ranking. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1386–1393 (2014)
16. Chen, Q., Huang, J., Feris, R., Brown, LM., Dong, J.: Deep domain adaptation for describing people based on fine-grained clothing attributes. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5315–5324 (2015)
17. Mairal, J., Koniusz, P., Harchaoui, Z., Schmid, C.: Convolutional kernel networks. In: International Conference on Neural Information Processing Systems. MIT Press, pp. 2627–2635 (2014)
18. Paulin, M., Douze, M., Harchaoui, Z., Mairal, J., Perronin, F.: Local convolutional features with unsupervised training for image retrieval. In: IEEE International Conference on Computer Vision, pp. 91–99 (2015)
19. Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K.: Visual genome: connecting language and vision using crowdsourced dense image annotations. Int. J. Comput. Vis. **123**(1), 32–73 (2017)
20. Kalantidis, Y., Mellina, C., Osindero, S.: Cross-dimensional weighting for aggregated deep convolutional features. In: Proceedings of the European Conference on Computer Vision. Workshops, pp. 685–701 (2016)

21. Tolias, G., Sicre, R., Jegou, H.: Particular object retrieval with integral maxpooling of CNN activations. In: Proceedings of the International Conference on Learning Representations, pp. 1–12 (2016)
22. Jegou, H., Zisserman, A.: Triangulation embedding and democratic aggregation for image search. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3310–3317 (2014)
23. Razavian, A.S., Sullivan, J., Maki, A., Carlsson, S.: Visual instance retrieval with deep convolutional networks (2014). arXiv:1412.6574