VLAD Is not Necessary for CNN

Dan Yu(✉) and Xiao-Jun Wu

School of IoT Engineering, Jiangnan University,
1800 Lihu Avenue, Wuxi, China
xiaojun_wu_jnu@163.com, wu_xiaojun@jiangnan.edu.cn

Abstract. Global convolutional neural networks (CNNs) activations lack geometric invariance, and in order to address this problem, Gong et al. proposed multi-scale orderless pooling(MOP-CNN), which extracts CNN activations for local patches at multiple scale levels, and performs orderless VLAD pooling to extract features. However, we find that this method can improve the performance mainly because it extracts global and local representation simultaneously, and VLAD pooling is not necessary as the representations extracted by CNN is good enough for classification. In this paper, we propose a new method to extract multi-scale features of CNNs, leading to a new structure of deep learning. The method extracts CNN representations for local patches at multiple scale levels, then concatenates all the representations at each level separately, finally, concatenates the results of all levels. The CNN is trained on the ImageNet dataset to extract features and it is then transferred to other datasets. The experimental results obtained on the databases MITIndoor and Caltech-101 show that the performance of our proposed method is superior to the MOP-CNN.

Keywords: CNN · Multi-scale · Deep learning · VLAD · Transfer learning

1 Introduction

Image classification [1–5] is one of the most important research tasks in computer vision and pattern recognition. To choose the right features plays the key role in a recognition system. There are many feature descriptors such as SIFT [6] and HOG [7], but they need to be designed by handcraft carefully, which is time-consuming and may not get the best feature sometimes. Many researches show that the features of the best performing recognition models are learned unsupervisedly from raw data.

Recently, deep convolutional neural networks (CNNs) have been considered as a powerful class of models for image recognition problems [8–11]. The feature representation learned by these networks achieves state-of-the-art performance not only on the task for which the network was trained, but also on various other classification tasks. A lot of recent works [12–14] showed that the feature representation trained on a large dataset can be successfully transferred to other visual tasks. For example: classification on Catech-101 [15], Catech-256 [5]; scene recognition on the Pascal VOC 2007 and 2012 [12] databases and so on.

However, global CNN activations lack geometric invariance, which limit their performance for the task of high variable scenes. Gong et al. [16] proposed a simple

G. Hua and H. Jégou (Eds.): ECCV 2016 Workshops, Part III, LNCS 9915, pp. 492–499, 2016.
DOI: 10.1007/978-3-319-49409-8_41

scheme called multi-scale orderless pooling CNN (MOP-CNN) to solve this problem, which combining activations extracted at multiple local image windows. The main idea of MOP-CNN is extracting features from the local patches via CNN at multiple scales, then adopting Vectors of Locally Aggregated Descriptors (VLAD) [17, 18] to encode those local features for each level separately, finally, concatenating the encoded features for all levels.

It is well known that the feature representation of CNN is very good, so is the VLAD really necessary? To explore this question, in this paper, we propose a method of MOP-CNN without the VLAD encoding. First, we extract local features via CNN at multiple scales, then we concatenate all the features at each level and PCA is adopted to reduce the dimensions of the concatenated features. Finally, we concatenate the features after PCA for all levels. We compare our proposed method with MOP-CNN on three datasets MITIndoor and Caltech-101 and evaluate their performances in accuracy and efficiency using strategy of transfer learning.

The rest of the paper is organized as follows. In Sect. 2, we introduce the proposed method in detail. Section 3 shows the experimental and compared results on the datasets MITIndoor and Caltech-101 respectively. We conclude the paper in Sect. 4.

2 The Proposed Method

We take the activation for the entire 256 × 256 image as the feature representation of the first level. For the second level, we extract activations for all 128 × 128 patches sampled with a stride of 32 pixels. Then we simply concatenate the activations for all

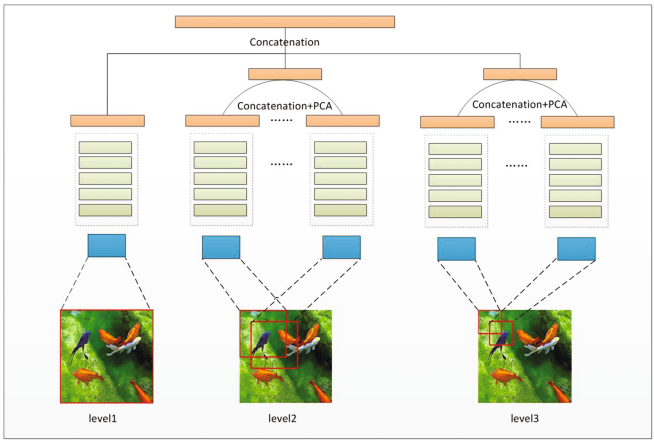

Fig. 1. Illustration of how the proposed method extracts features from an image through multi-scale concatenation for CNN activations. There are also three levels in our method: Level 1 extracts the 4096-dimension feature of the last connected layer of CNN for the entire 256 × 256 image. Level 2 extracts 4096-dimension representation for each 128 × 128 patch and concatenates all representations of all patches from the image, which is then reduced to 4096-dimension via PCA. Level 3 formed in the same way as level 2 but replaces the patch size 128 × 128 with 64 × 64. Finally, we concatenate all the features of three levels.

patches, which results in quite high dimensional vector, so we use PCA to reduce them to 4096, finally, the reduced feature vectors are normalized as the final feature representation of the second level. The third level is the same as the second level but replacing the patches size 128×128 with 64×64, which can extract more local information intuitively (but we found it does not work well, which we will discuss in Sect. 3). Finally, we concatenate the original 4096-dimensional feature representation from the first level and the two PCA-reduced 4096-dimensional feature representations from the second and third levels to form the final feature representation of an image (shown as Fig. 1).

A direct transfer learning strategy is adopted for visual classification. The CNN is trained on the ImageNet to extract features and it is then transferred to other datasets. In order to indicate the ability to learn rich image representations of CNN, we reuse layers trained on the ImageNet without fine-tuning. The main idea is shown in Fig. 2. A CNN representation trained on the Imagenet dataset used on other dataset is a standard practice now, but it is a transfer procedure.

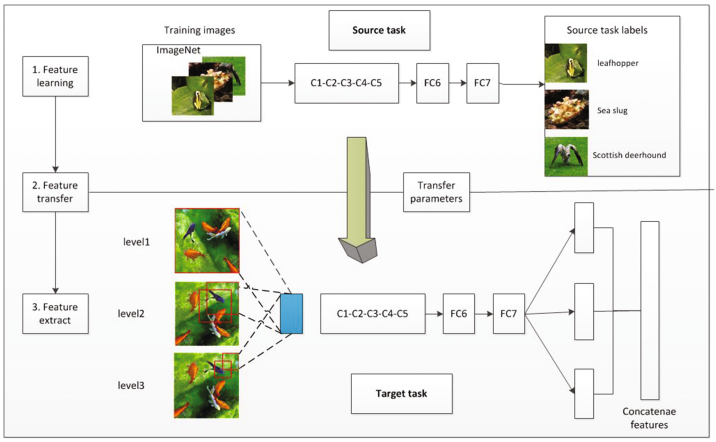

Fig. 2. Transferring parameters of a CNN. The network is trained on the source task (ImageNet classification), then the parameters of the internal layers of the network (C1-FC7) are transferred to the target tasks. In this paper, we reuse the parameters directly without fine-tuning.

3 Experimental Results

In this section, we evaluate and compare our proposed method with MOP-CNN on two datasets: MITIndoor and Caltech-101. We also discuss the performance of each level, which means the global and local information of an image.

3.1 DataSets

MITIndoor [20] contains 67 categories, and a total of 15620 images. There is a standard training/test split, which contains 80 training and 20 test images per category.

Caltech-101 [21] contains 101 categories, and about 40 to 800 images per category, most categories have about 50 images. We follow the procedure of [22] and randomly select 5,10,15,20,25 images per class for training and test on up to 20 images per class, repeat 5 times and report the average of the per-class accuracy.

3.2 Results

In all the experiments of this paper, we adopt the SVM [23–25] implementation from the libsvm [26, 27] as the classifier.

The results on MITIndoor is shown in Table 1. From Table 1, one can see that simply concatenating the features of all patches is better than VLAD pooling, which implies that we can extract pretty good features for classification just via CNN and without VLAD encoding. And the training time and test time of our proposed method are shorter than that of VLAD encoding. One can also see that the concatenation of level 1 and level 2 achieves best recognition accuracy, which may because level 1 can extract the global feature and level 2 can extract the local feature, and concatenating level 1 and level 2 can obtain the local and global information simultaneously to improve the recognition accuracy. That means that the multi-scale information is useful to improve the performance of CNN. However, concatenating all the three scale levels is not very good, it may because the patch size of level 3 is too small, which could not extract the main discriminative information and may introduce some noises.

Table 1. Performance on MITIndoor

Pooling method	Scale	Training time (1.0e+04 *) (s)	Test time (s)	Acc (%)
VLAD pooling	level1	1.06	1.36	59.68
	level2	1.75	2.52	54.80
	level3	2.51	3.95	51.88
	level1+level2	2.19	3.38	63.29
	level1+level3	2.98	4.86	63.81
	level2+level3	3.65	5.88	57.88
	level1+level2+level3 (MOP-CNN)	4.12	6.86	64.34
Concatenation+PCA	level1	0.52	0.95	59.68
	level2	1.17	2.10	58.41
	level3	1.96	3.49	52.85
	level1+level2 (Our method)	1.69	2.99	**64.34**
	level1+level3	2.47	4.40	63.44
	level2+level3	3.11	5.51	58.33
	level1+level2+level3	3.68	6.37	63.81

We implement the experiments of MOP-CNN using the same experimental conditions as Gong et al. First, we extract multi-scale features on different patches size via CNN, then use VLAD to encode the features, the parameters of VLAD is the same as

Gong et al. But the results is worse than that reported in the MOP-CNN paper from Gong et al., which may come from two implementation details: one possible reason is that we use the CNN trained on the ImageNet directly without fine-tuning on the target datasets. However, fine-tuning was not reported in [16] explicitly. Another reason may be from different implementations of SVM classifier. We adopt the SVM implementation from the libsvm [26, 27] rather than the linear SVM implementation from the INRIA JSGD package on [16].

Table 2 shows the results on Caltech-101 of 20 images per class for training and up to 20 images per class for test. Figure 3 shows the results of different training images. From Table 2, we can see that the trends are consistent with those on MITIndoor, which implies that our proposed method is superior to MOP-CNN, which means VLAD is not necessary. There is one interesting difference from Table 1, the concatenation of level 1 and level 2 performs much better than level 1 or level 2 alone on MITIndoor, while the advantage is not very significant on Caltech-101. The possible reason is that indoor scenes are better described by the concatenation of local and global discriminative information. From Fig. 3 we can see that the performance increases as more training images are used, and our method is better than MOP-CNN no matter how training images are used.

Table 2. Performance on Caltech-101

Pooling method	Scale	Training time (1.0e+04 *) (s)	Test time (s)	Acc (%)
VLAD pooling	level1	0.26	1.17	86.44
	level2	0.48	2.25	69.42
	level3	0.78	3.75	50.53
	level1+level2	0.65	3.08	85.36
	level1+level3	0.95	4.53	85.06
	level2+level3	1.17	5.61	68.21
	level1+level2+level3 (MOP-CNN)	1.34	6.44	83.98
Concatenation +PCA	level1	0.19	0.86	86.44
	level2	0.41	1.94	82.71
	level3	0.70	3.39	64.88
	level1+level2 (Our method)	0.58	2.77	**88.31**
	level1+level3	0.87	4.27	86.68
	level2+level3	1.09	5.30	78.78
	level1+level2+level3	1.26	6.12	87.06

From the experimental results on the two datasets, we can conclude that: (a) The features extracted via CNN is good enough for the recognition tasks and the simple concatenation of the features of level 1 and level 2 is better than the features via VLAD encoding no matter in performance or time consumption, and no matter how training

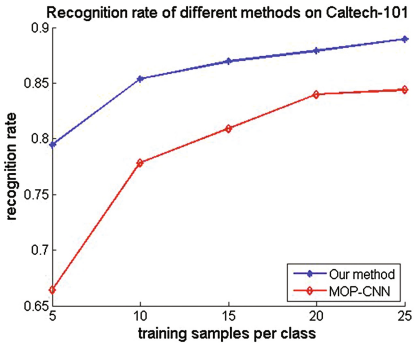

Fig. 3. Recognition rates of different methods on Caltech-101

images are used, which means VLAD is not necessary. (b) The concatenation of level 1 and level 2 is superior to level 1 or level 2 alone shows that the multi-scale information is useful to improve the performance of CNN, while the features of level 3 is not as good as level 1 and level 2, that probably because the patch size of level 3 is too small to capture discriminative information and may introduce noise. (c) The contribution of the local information varies from datasets, indoor scenes are better described by local patches that have highly distinctive appearance but can vary greatly in terms of location.

4 Conclusion

In this paper, we propose a new simple method to extract multi-scale feature representation of CNN, which concatenates the features of all patches on each level simply, rather than using VLAD encoding. The experimental results on two datasets: MITIndoor and Caltech-101 show that the features extracted by CNN are good enough for classification tasks and VLAD encoding is not necessary. From the experimental results, we can also learn that the multi-scale information is helpful but the patch size is important for the extraction of local information, while it may not be helpful if the patch is too small. Furthermore, we can see that the contribution of the local information is specific to datasets depending on the visual content of images.

In this paper, we only discuss the classification task, and there are many other tasks in computer vision and pattern recognition. In the future, we will study the influence of the multi-scale features to other tasks, such as detection task, localization task and so on.

Acknowledgments. The paper is supported by the National Natural Science Foundation of China (Grant No. 61373055, 61672265), Industry Project of Provincial Department of Education of Jiangsu Province (Grant No. JH10-28), and Industry oriented project of Jiangsu Provincial Department of Technology (Grant No. BY2012059).

References

1. Chan, T.H., Jia, K., Gao, S., Lu, J., Zeng, Z., Ma, Y.P.: A simple deep learning baseline for image classification? arXiv preprint arXiv:1404.3606 (2014)
2. Kavukcuoglu, K., Sermanet, P., Boureau, Y.L., Gregor, K., Mathieu, M., Cun, Y. L.: Learning convolutional feature hierarchies for visual recognition. In: Advances in Neural Information Processing Systems (NIPS), pp. 1090–1098 (2010)
3. Hillel, A.B., Weinshall, D.: Subordinate class recognition using relational object models. In: Advances in Neural Information Processing Systems (NIPS) (2007)
4. Berg, T., Belhumeur, P.: POOF: part-based one-vs.-one features for fine-grained categorization, face verification, and attribute estimation. In: CVPR, pp. 955–962 (2013)
5. Donahue, J., Jia, Y., Vinyals, O., Hoffman, J., Zhang, N., Tzeng, E., Darrell, T.: Decaf: a deep convolutional activation feature for generic visual recognition. arXiv preprint arXiv:1310.1531 (2013)
6. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vis. (IJCV) **60**, 91–110 (2004)
7. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: CVPR, pp. 886–893 (2005)
8. Huang, F.J., LeCun, Y.: Large-scale learning with SVM and convolutional netw for generic object recognition. In: CVPR (2006)
9. Krizhevsky, A., Sutskever, I., Hinton, G. E.: Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems (NIPS), pp. 1097–1105 (2012)
10. Chen, Y.N., Han, C.C., Wang, C.T., Jeng, B.S., Fan, K.C.: The application of a convolution neural network on face and license plate detection. In: ICPR, pp. 552–555 (2006)
11. Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., Fei-Fei, L.: Large-scale video classification with convolutional neural networks. In: CVPR, pp. 1725–1732 (2014)
12. Oquab, M., Bottou, L., Laptev, I., Sivic, J.: Learning and transferring mid-level image representations using convolutional neural networks. In: CVPR, pp. 1717–1724 (2014)
13. Sermanet, P., Eigen, D., Zhang, X., Mathieu, M., Fergus, R., LeCun, Y.: Overfeat: integrated recognition, localization and detection using convolutional networks. arXiv preprint arXiv:1312.6229 (2013)
14. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
15. Zeiler, M.D., Fergus, R.: Visualizing and understanding convolutional networks. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8689, pp. 818–833. Springer, Heidelberg (2014). doi:10.1007/978-3-319-10590-1_53
16. Gong, Y., Wang, L., Guo, R., Lazebnik, S.: Multi-scale orderless pooling of deep convolutional activation features. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8695, pp. 392–407. Springer, Heidelberg (2014). doi:10.1007/978-3-319-10584-0_26
17. Jégou, H., Douze, M., Schmid, C., Pérez, P.: Aggregating local descriptors into a compact image representation. In: CVPR, pp. 3304–3311 (2010)
18. Perronnin, F., Dance, C.: Fisher kernels on visual vocabularies for image categorization. In: CVPR, pp. 1–8 (2007)
19. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems (NIPS), pp. 1097–1105 (2012)
20. Quattoni, A., Torralba, A.: Recognizing indoor scenes. In: CVPR, pp. 413–420 (2009)

21. Fei-Fei, L., Fergus, R., Perona, P.: Learning generative visual models from few training examples: an incremental bayesian approach tested on 101 object categories. Comput. Vis. Image Underst. **106**, 59–70 (2007)
22. Fei-Fei, L., Fergus, R., Perona, P.: One-shot learning of object categories. IEEE Trans. Pattern Anal. Mach. Intell. **28**, 594–611 (2006)
23. Burges, C.J.: A tutorial on support vector machines for pattern recognition. Data Min. Knowl. Discov. **2**, 121–167 (1998)
24. Cristianini, N., Shawe-Taylor, J.: An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods. Cambridge University Press, Cambridge (2000)
25. Smola, A.J., Schölkopf, B.: A tutorial on support vector regression. Stat. Comput. **14**, 199–222 (2004)
26. Hsu, C.W., Chang, C.C., Lin, C.J.: A practical guide to support vector classification (2003)
27. Chang, C.C., Lin, C.J.: LIBSVM: a library for support vector machines. ACM Trans. Intell. Syst. Technol. (TIST) (2011)