15th IAPR International Conference on Machine Vision Applications (MVA)
Nagoya University, Nagoya, Japan, May 8-12, 2017.

Deep Visual Words: Improved Fisher Vector for Image Classification

Ali Diba
PSI-ESAT
adiba@esat.kuleuven.be

Ali Mohammad Pazandeh
Sharif Tech
pazandeh@ee.sharif.edu

Luc Van Gool
PSI-ESAT / ETHZ
vangool@vision.ee.ethz.ch

Abstract

Image classification has been revolutionized by deep convolutional neural networks. Using previous state-of-the-art classification methods like Fisher vector encoding in combination with deep CNNs has been shown to be promising. Motivated by the recent work on dense CNN features to extract Fisher encoding(FV-CNN), we present a scheme to discover better visual words with CNNs, to obtain improved Fisher vector features. Our method (Deep Visual Words-DVW) learns semantic visual clusters per each category, by iteratively learning and refining groups of visual patches. DVW represents an efficient feature space embedding to capture the discriminative potential between meaningful visual clusters. We evaluate our approach on popular datasets in object, scene and action classification and outperformed the state-of-the-art: scene classification MIT indoor, object categorization PASCAL VOC 2007 and Stanford40 human actions.

1 Introduction

The discovery of effective image representations plays an important role in achieving promising visual recognition performance. Deep convolutional neural networks [1] can be helpful for this purpose. Recent works [2, 3, 4, 5, 6] and the results thereof prove this point. Previously, Bag-of-Word models have been shown to extract global image features that aggregate local image descriptors into better representations. These are robust to variations in visual appearance like scaling, rotation and translation. The most significant improvement in these models is the Fisher kernel [7, 8].

To exploit CNN activation as a generic image feature, we can extract information from different layers of the CNN, and in particular from the fully connected layers where the image has been pushed into a pre-trained CNN model. This setting may still fail to sufficiently mind the visual variations though. Cimpoi *et al.* [9] proposed FV-CNN, a pooling method to overcome these issues. FV-CNN uses CNN features extracted from dense image patches and encodes them by Fisher vectors. This work is similar in flavour to other methods like [10, 11], that combine pooling strategies with CNN activations. These works have shown that using CNN features for image patches or parts and encoding them using methods like Fisher vectors can perform well.

In this paper, we propose a simple and more efficient method to automatically discover visual category parts. We also learn and fine-tune a deep CNN feature extractor for each part concept. We re-design discriminative patch mining methods [4, 12] and use deep CNN feature learning as a module in the loop. We demonstrate that this 1) yields more effective representations for visual words and 2) configures meaningful discriminative patch clusters as visual words. Utilizing these new visual words in Fisher vector encoding yields improvement for both.

Selecting distinctive parts and reformulating their representation are the main goals that we aim for. Our iterative pipeline can reconfigure groups of visual parts and refine their feature learning process in each iteration. Our experiments prove that these new deep visual words can be effective when combined with a Fisher vector representation and that they yield promising results for scene classification (MIT indoor 67 [21]) and object classification (PASCAL VOC 2007 [22]) and human actions (Stanford40 [24]).

The rest of the paper is organized as follows. Related work is discussed in section 2. Section3 describes our proposed framework. Section 4 evaluates our method and compares the results with the state-of-the-art. Section 5 concludes the paper.

2 Related Works

There has been quite some works related to discriminative image parts learning and mid-level visual recognition [13, 14, 15, 9]. These contributions share the use of some part discovery modules and an encoding method to efficiently represent the parts. Image classification works better by encoding visual words. The main difference between the existing work and ours is that we involve feature learning throughout the entire pipeline.

Juneja *et al.* [15] proposed a simple and effective method to evaluate the distinctiveness of visual clusters, which is based on the ranking of entropy. Singh *et al.* [13] probe visual part mining in a weakly supervised manner, using image level labels. Diba et al. [23] has introduced an iterative CNN training for mining discriminative patches to recognize actions and human attributes.

Cimpoi *et al.*s work [9] is similar in vein as our work, in that they use Fisher vector encoding of local CNN activations, for image classification and texture recognition. Our work is different in the way we apply a method to discover more effective visual patterns and words to Fisher encoding.

3 Method

In this section we discuss our pipeline for image classification. After presenting the overview and the motivation behind the work, we introduce the proposed feature extraction. In the final part the encoding method is explicated.

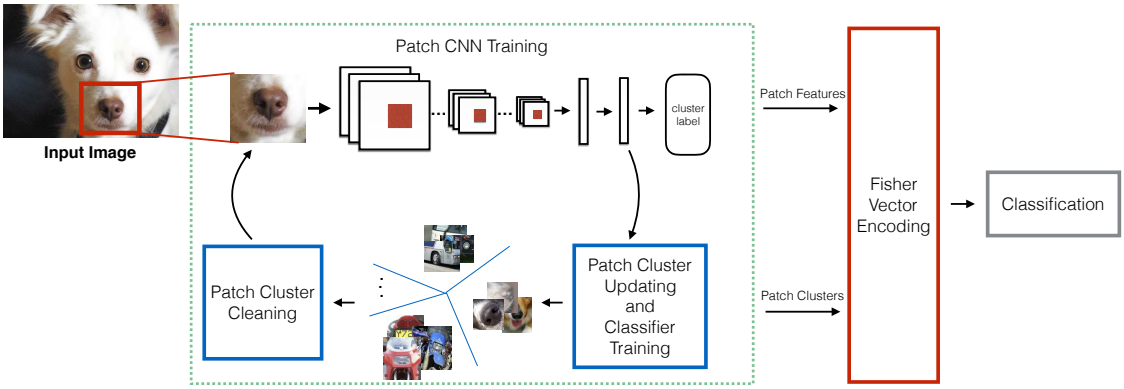

Figure 1. Flow of our pipeline: Iterative learning CNN for ditinctive patch clusters, re-clustering them and cleaning outliers from clusters. After configuring patch clusters and their CNN network, classification images by encoding local features using fisher encoding and new CNN features is done.

3.1 Approach overview

The main idea of the approach is to exploit new visual words and efficient feature embedding for them to obtain powerful local features aggregation by fisher vector encoding. Similarly, [9] extracts dense local CNN features and encode them by fisher vector, but there is no procedure to mining visual words or refining their local feature representation.

We introduce an iterative pipeline to extract more discriminant features out of densely distributed patches. To encode the extracted patch features we first use the common Fisher vector encoding method [8]. To improve the performance of recognition, we propose a modification in Fisher vector encoding method. It aggregates the features extracted from dense patches to produce a global feature for the whole image. The overview of the proposed pipeline is shown in Fig. 1.

The features which are extracted from a convolutional neural network trained on image patch - class label pairs, still show a weak feature embedding. This due to a mismatch between the nature of the input patches and their output labels. In addition, the classification of the patches is only weakly supervised, because patch labels do not come with the training set. To improve the feature embedding for patches we need a module that simultaneously extracts discriminant features and predicts the subclass labels of the patches. We discuss the details of the proposed module in section 3.2.

As we discuss in detail in section 3.3, in addition to achieving a powerful feature embedding, the proposed iterative procedure yields discriminant subclass clusters. Hence, we use the extracted discriminant clusters instead of the GMM clustering procedure of standard Fisher vector encoding.

3.2 Discriminant Feature Extraction

In this section, we introduce our iterative discriminant feature extraction and clustering method, and the processing blocks in the corresponding pipeline.

To achieve the aforementioned desired properties of the module powerful embedding and producing discriminant clusters we need carry out the following tasks in an iterative manner. Firstly, we extract features from the image patch using the CNN trained on cluster labels. Secondly, we update the cluster labels to maximize the discrimination between the clusters. Thirdly and finally, we clean the clusters to render them as rich as possible, because we believe removing outlier samples of clusters and retraining CNN with cleaner clusters allows enriched representation. Details of these processing blocks follow next.

3.2.1 Patch Convolutional Neural Network

The CNN block is the basis of the proposed iterative procedure. It is trained to extract discriminant features from the image patches. In the training phase the inputs of the model are the extracted image patches, and the corresponding labels are the cluster labels produced by the previous iteration. After the network is trained, the output of the network for an image patch is the final $fc7$ layer feature.

3.2.2 Patch Cluster Updating and Classifier Training

The cluster updating module is based on subdividing each class into a specific number of subclasses (clusters). The goal is to find clusters that discriminate a subclass against other subclasses, as well as against other classes at the parent level. To do so we apply the Mid-Level Deep Pattern Mining (MDPM) algorithm [4] to the features of the previous stage.

The classifier training procedure in this block is done by training an LDA classifier on the extracted features as input and the computed clusters as labels.

3.2.3 Cluster Cleaning

Cleaning the clusters by picking out weakly discriminant patch instances and putting them in a negative class, renders clusters more discriminant and subsequently yields a more powerful feature embedding. This is achieved by thresholding the output score of the LDA classifiers of the previous phase.

After an adequate number of iterations, the discriminant patch features and clusters are available for encoding.

Method	mAP(%)
MDPM [4]	75.2
Deep Filter Banks [9]	76.4
SCFVC *et al.* [20]	76.9
CNN (VGG-M net)	72.2
Ours_Fisher	77.4
Ours_Modified_Fisher	**79.2**

Table 1. Mean average precision on the PASCAL VOC 2007 object dataset and comparison with previous methods. The results are reported on the test set of PASCAL VOC. The first row shows the baseline of the work.

3.3 Feature Encoding

The last stage before training a classifier and actually applying it for recognition is to encode the extracted features from the patches, in order to aggregate them into a global image feature. The success of previous work that used Fisher vector encoding [7] suggests that scheme for our dense features encoding. As the results of the experimental section show, the proposed iterative process adds power to the feature embedding. As a consequence, the features conducted from Fisher vector encoding become more discriminant and improve the performance of image classification.
On the other hand, Fisher vector encoding is based on a visual dictionary learning method using Gaussian Mixture Models (GMM), which does not sufficiently guarantee efficient discriminating performance by the output clusters of the dictionary, despite the effort of doing so. This weakness motivates us to use the discriminative output clusters of the iterative procedure as the visual dictionary clusters instead of those from GMM learning. By so doing, we get a significant improvement in recognition accuracy. This testifies to the power of the output clusters when it comes to discriminating between themselves. The cause of this clustering process being more powerful than others lies in the joint feature learning and cluster cleaning.

The final stage of the method classifies the encoded global features. As the output feature of the Fisher vector encoding is linear separable we use a linear multi-class SVM to make prediction in the testing phase .

4 Experiments

We have tested our proposed framework on two classification tasks, i.e. object and scene and action recognition. The following sections expound the details of the implementation, datasets, and results.

4.1 Experimental Setup

This section introduces the experimental setup of the convolutional neural network and of the encoding based on Fisher vectors.

4.1.1 Patch Convolutional Neural Network

The network has been trained using the caffe CNN training package [17] with back-propagation. The used CNN model is VGG-M [18] which performs better than regular models like [2] with the same cost. We use the weights of the network trained on the ImageNet dataset [19] as initial weights and fine-tune our networks on the object, scene and action datasets. We set the learning rate of CNN training to 0.0001, and the batch size to 100.
The patches are densely sampled from the input images at 3 different scales (128*128, 160*160, and 192*192 patches from a resized image with a stride of 16).

4.1.2 Fisher Vector Encoding

The learned Fisher dictionaries have 64 Gaussian components, resulting in a 52K-dimensional descriptor. To create the dictionary of the modified Fisher encoding, we use all instances with a specific cluster label and calculate their mean and variance for each cluster.

4.2 Dataset:

Object: The evaluation dataset used for object class recognition is the PASVAL-VOC 2007 dataset. This dataset contains 9963 images containing 24,640 annotated objects for 20 different classes, split into training, validation, and testing subsets.
Scene: For evaluating our method on scene recognition, we use the MIT Indoor Scene dataset. The dataset consists of 15620 instances of 67 indoor categories. The dataset is split into two training and testing subsets with equal numbers of images.
Action: The Stanford40 action dataset contains total of 9532 images and 40 classes of actions, split into train set containing 4000, and test set containing 5532 instances.

4.3 Object Recognition

We report the resulting accuracy of object recognition in table 1 and compare those results with the reported mean average precision of [4, 9, 20]. As the Deep filter banks, MDPM and CNN methods are each, parts of our proposed method they are supposed as the baselines of the method. Ours_Fisher and Ours_Modified_Fisher report the results of our approach, using the standard Fisher vector encoding and the proposed modified Fisher vector encoding. As it can inferred from the results we perform state-of-the-art in object recognition task on the PASCAL-VOC 2007 dataset.

4.4 Scene and Aaction Recognition

The classification results of the proposed methods on the MIT Indoor Scene dataset are reported in Table 2. We contrast our results with the results of [4, 9, 20, 15]. Same as the reported results of object recognition in table 1, the Deep filter banks, MDPM and CNN methods could be considered as the baselines of our method. The other rows on top of the table report the average precision of other recent works on the dataset, and the two bottom rows are the results of the proposed method. As the reported results show we achieve state-of-the-art on the MIT indoor dataset in scene recognition task same as in object recognition. We have

Method	MIT 67	Stanfrod40
Juneja *et al.* [15]	63.18	–
ObjectBank [25]	–	32.5
EPM [26]	–	40.7
Dorsch *et al.* [14]	64.03	–
SCFVC *et al.* [20]	68.2	–
MDPM [4]	69.7	46.8
Deep Filter Banks [9]	74.2	–
CNN (VGG-M net)	62.5	45.9
Ours_Fisher	75.9	49.8
Ours_Modified_Fisher	**77.4**	53.2

Table 2. Mean average precision on the MIT indoor scene and Stanford40 action dataset and comparison with previous methods.

a similar for recognizing actions. The evaluation of our approaches for actions can be found at Table 2. The results present an outperforming over the previous methods in this human action dataset.

5 Conclusion

Strong description of local patches and meaningful visual words will improve image classification performance and provide more semantic representation for this task. In this work, we propose an iterative pipeline to learn visual clusters and also CNN feature embedding jointly. The learned discriminative visual clusters and their new CNN feature extractor benefit fisher vector encoding to outperform state-of-the-art methods in image classification which we prove it by our results on scene and object categorization task.

References

[1] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, 1998.

[2] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012.

[3] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *CVPR*, 2014,.

[4] Yao Li, Lingqiao Liu, Chunhua Shen, and Anton van den Hengel, "Mid-level deep pattern mining," in *Computer Vision and Pattern Recognition (CVPR), .* , 2015,.

[5] Bolei Zhou, Agata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva, "Learning deep features for scene recognition using places database," in *Advances in neural information processing systems*, 2014,.

[6] Jonathan Long, Evan Shelhamer, and Trevor Darrell, "Fully convolutional networks for semantic segmentation," in *CVPR*, 2015, .

[7] Florent Perronnin, Jorge Sánchez, and Thomas Mensink, "Improving the fisher kernel for large-scale image classification," in *Computer Vision–ECCV 2010*, .

[8] Florent Perronnin and Christopher Dance, "Fisher kernels on visual vocabularies for image categorization," in *Computer Vision and Pattern Recognition,. CVPR* . 2007.

[9] Mircea Cimpoi, Subhransu Maji, and Andrea Vedaldi, "Deep filter banks for texture recognition and segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

[10] Yunchao Gong, Liwei Wang, Ruiqi Guo, and Svetlana Lazebnik, "Multi-scale orderless pooling of deep convolutional activation features," in *Computer Vision–ECCV 2014*, 2014.

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," in *Computer Vision–ECCV 2014*, 2014.

[12] Xinggang Wang, Baoyuan Wang, Xiang Bai, Wenyu Liu, and Zhuowen Tu, "Max-margin multiple-instance dictionary learning," in *ICML*, 2013.

[13] Saurabh Singh, Abhinav Gupta, and Alexei Efros, "Unsupervised discovery of mid-level discriminative patches," *Computer Vision–ECCV 2012*, 2012.

[14] Carl Doersch, Abhinav Gupta, and Alexei A. Efros, "Mid-level visual element discovery as discriminative mode seeking," in *NIPS*, 2013.

[15] Mayank Juneja, Andrea Vedaldi, CV Jawahar, and Andrew Zisserman, "Blocks that shout: Distinctive parts for scene classification," in *CVPR*, 2013,.

[16] Sobhan Naderi Parizi, Andrea Vedaldi, Andrew Zisserman, and Pedro Felzenszwalb, "Automatic discovery and optimization of parts for image classification," *arXiv preprint arXiv:1412.6598*, 2014.

[17] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell, "Caffe: Convolutional architecture for fast feature embedding," in *ACM MultiMedia*. 2014,.

[18] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman, "Return of the devil in the details: Delving deep into convolutional nets," in *BMVC*, 2014.

[19] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Computer Vision and Pattern Recognition, 2009.* .

[20] Lingqiao Liu, Chunhua Shen, Lei Wang, Anton van den Hengel, and Chao Wang, "Encoding high dimensional local features by sparse coding based fisher vectors," in *Advances in Neural Information Processing Systems*, 2014.

[21] Quattoni, Ariadna and Torralba, Antonio, "Recognizing indoor scenes," in *CVPR*, 2009.

[22] Everingham, Mark and Van Gool, Luc and Williams, Christopher KI and Winn, John and Zisserman, Andrew "The pascal visual object classes (voc) challenge," in *International journal of computer vision*, 2010.

[23] Diba, Ali et.al "Deepcamp: Deep convolutional action & attribute mid-level patterns," in CVPR, 2016.

[24] Yao, Bangpeng and Jiang, Xiaoye and Khosla, Aditya and Lin, Andy Lai and Guibas, Leonidas and Fei-Fei, Li "Human action recognition by learning bases of action attributes and parts," in *ICCV*, 2011.

[25] Li, Li-Jia and Su, Hao and Fei-Fei, Li and Xing, Eric P "Object bank: A high-level image representation for scene classification & semantic feature sparsification," in *NIPS*, 2010.

[26] Sharma, Gaurav and Jurie, Frédéric and Schmid, Cordelia "Expanded parts model for human attribute and action recognition in still images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013.