Bag of Surrogate Parts: one inherent feature of deep CNNs

Yanming Guo
y.guo@liacs.leidenuniv.nl
Michael S. Lew
mlew@liacs.nl

LIACS Media Lab, Leiden University
Niels Bohrweg 1, Leiden
the Netherlands

Abstract

Convolutional Neural Networks (CNNs) have achieved promising performance in image classification tasks. In this paper, we develop a new feature from convolutional layers, called Bag of Surrogate Parts (BoSP), and its spatial variant, Spatial BoSP (S-BoSP). Specifically, we take the feature maps in convolutional layers as surrogate parts, and densely sample and assign the regions in input images to these surrogate parts by observing the activation values. To better handle the objects with different scales and deformations, and make more comprehensive predictions, we further propose a scale pooling technique for assigning the features, and global constrained augmentation for the final prediction. Compared with most existing methods that also utilize the activations from convolutional layers, the proposed method is efficient, has no tuning parameters, and could generate low-dimensional, highly discriminative features. The experiments on generic object, fine-grained object and scene datasets indicate that the proposed feature can not only produce superior results to fully-connected layer based features, but also get comparable, or in some cases considerably better performance than the state-of-the-art.

1 Introduction

In recent years, convolutional neural networks (CNNs) have been widely employed to address the image classification challenge and have produced leading performance on various benchmark datasets [11]. It has been verified that, the network pretrained on a large and diverse dataset, such as ImageNet [5], can act as a universal model and be transferred to other visual recognition tasks [6, 10, 20, 24].

Generally, a CNN consists of three types of layers: convolutional layers, pooling layers and fully-connected layers. In contrast to most of the research that take the activations from fully-connected layers as the image representation [1, 34], we derive a new feature from convolutional layers, called Bag of Surrogate Parts (BoSP), by borrowing the idea of the well-known Bag-of-Words (BoW) scheme [28]. The main idea is: we take the feature maps in convolutional layers as surrogate parts, and take the activation values on the feature maps as assignment strengths for these surrogate parts. Since each unit in the feature map corresponds to one receptive field, this operation performs like densely sampling and assigning regions from the input image. The final feature is generated by concentrating the assignment strengths of the surrogate parts. In contrast to the conventional

Pages 96.1-96.12
DOI: https://dx.doi.org/10.5244/C.30.96

BoW scheme, BoSP deals with regions, rather than keypoints, and attempts to assign the regions to surrogate parts, which makes it more semantically meaningful. Besides, these surrogate parts have been inherently defined by the architecture, i.e. the feature maps, thus eliminates the time-consuming process of generating visual dictionaries. Our main contribution of BoSP is different from prior research [3, 4, 17, 18, 31] in that it considers and utilizes the convolutional features in a novel way, which incorporates the statistical information of the image and does not need to be tuned.

On top of BoSP, there are three other contributions of this paper:

(1) We propose a variant of BoSP, i.e. Spatial-BoSP(S-BoSP), to incorporate more spatial information, by dividing the image into several regions and concentrating the BoSP inside each sub-region. This method has been widely adopted in BoW-based approaches [15, 31].

(2) We develop a scale pooling scheme for the assignment of the units. The scheme allows us to process receptive fields of various sizes, and can handle the deformation problem inside each region by utilizing the max pooling technique. Scale pooling improves the performance considerably without enlarging the feature dimension.

(3) We raise a global constrained augmentation approach to incorporate the predictions of global feature and augmented feature.

Along with the contributions, there are several potential advantages for the proposed BoSP/S-BoSP:

(1) High accuracy. BoSP/S-BoSP could not only achieve superior performance compared to the corresponding top-layer activation, they can also deliver comparable or better performance than the state-of-the-art. For example, the accuracy of S-BoSP on Caltech101 dataset is 93.99% without any data augmentation, already slightly better than the state-of-the-art, which is 93.42%. After incorporating the proposed global constrained augmentation, it further improves the state-of-the-art to 94.52%.

(2) No tuning. BoSP/S-BoSP are inherent features of the architecture, i.e. the number of surrogate parts is pre-defined by the structure and the assignment strengths for the surrogate parts can be directly achieved by observing the activation values. There is no tuning parameters for us to determine.

(3) Memory efficient. Compared to other features which are also derived from convolutional layers [17], or other encoding schemes which combine CNN features and the variants of BoW [31, 33], BoSP/S-BoSP are relatively low-dimensional (512-D/4608-D), so they are advantageous in large scale applications.

2 Related Work

BoW-based scheme in deep CNN structures. BoW methods have been widely used in computer vision systems and achieved state-of-the-art performance. Recently, several studies attempted to introduce the idea of BoW in the CNN structure. Most of the approaches utilized its variants, such as VLAD [13] and Fisher vector [21]. For example, Gong et al. [9] proposed a Multi-scale Orderless Pooling (MOP) scheme which applied VLAD to encode the fully-connected activations from multiple scales, and then concentrated them together. Similarly, Yoo et al. [33] also aggregated multi-scale top layer activations of CNN by using the Fisher kernel. Although these approaches produced encouraging results on several datasets, the generation of the feature dictionaries is a computationally expensive process and needs to be designed carefully. In contrast, the BoSP/S-BoSP in this paper are inherently generated by the architecture and we do not need to tune any parameters to obtain them. This avoids the time-consuming and sensitive process of dictionary learning.

Typical usage of convolutional features. The usage of convolutional features can be divided into two approaches. In the first approach, researchers utilize the variants of BoW scheme, such as VLAD and Fisher vector, to encode the convolutional features. For example, Ng et al. [18] employed VLAD encoding method to encode features from different convolution layers and demonstrated that the intermediate layers produce better results for image retrieval than the top layers. In contrast, Cimpoi et al. [4] and Wei et al. [31] took advantage of Fisher vector to encode the deep descriptors from intermediate layers, and also achieved promising performance on their tasks. In the second approach, researchers make use of the convolutional activations in a more straightforward way, by aggregating and compressing them to form the final representation. For instance, Liu et al. [17] took the feature maps as the indicator maps of parts, and aggregated the local features of each surrogate part as the image-level representation. Despite promising results, most of these schemes generate high-dimensional features and thus need additional computationally intensive post-processing, e.g. PCA compression or whitening. Our method can be seen as the combination of these two approaches. Similar to [17], we regard the feature maps as surrogate parts, but do not explicitly calculate and concentrate the features of these parts. Instead, we only aggregate their statistical strengths, which makes the feature dimension much lower.

3 Proposed method

In this part, we first describe our proposed BoSP/S-BoSP feature, and then introduce two approaches to further enhance the performance of the feature: scale pooling and global constrained augmentation.

3.1 Bag of Surrogate Parts Feature

The general CNN structure consists of multiple layers. When we feed in one image, it will convolve with multiple kernels to generate various feature maps. Each unit on the feature map corresponds to one receptive field on the input image, and the units in back feature maps correspond to larger receptive fields than the front ones.

With the intuition that larger receptive fields convey more semantic information, we extract BoSP from the pool$_5$ layer (i.e. the last pooling layer) of the VGG model [27]. The specific procedure is: we take the feature maps as surrogate parts and assume that the activation values represent the assignment strengths for these parts. Therefore, given the architecture, the number of the surrogate parts is inherently determined, as is the same with the number of feature maps. For each spatial unit, we can calculate its assignment strengths for the surrogate parts by observing its activation values. The one-by-one processing of these spatial units can be viewed as densely sampling and assigning regions of the input image. Finally, we sum the assignment strengths for the surrogate parts and form a vector accordingly, i.e. BoSP, whose length is the same with the number of the feature maps. The framework of the proposed BoSP feature is shown in Figure 1.

Specifically, suppose there are M surrogate parts, and for each image, we can densely sample n regions (for the pool$_5$ layer of VGG, $M = 512, n = 49$). The BoSP for this image

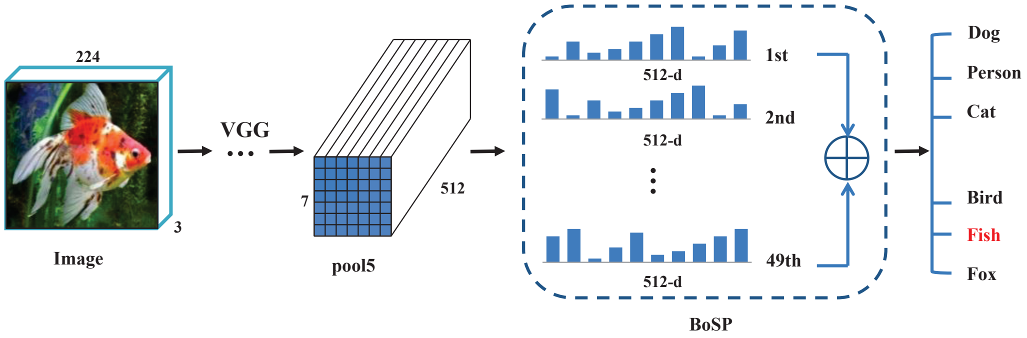

Figure 1: The framework to extract BoSP from the pool$_5$ layer of VGG. We can calculate the statistical histogram of the surrogate parts by observing the activation values.

can be represented as Eq.(1):

$$BoSP = \sum_{i=1}^{n} [P_1^i, P_2^i, \cdots, P_j^i, \cdots, P_M^i] \tag{1}$$

P_j^i is the assignment strength of region i on surrogate part j. We normalize the activations by dividing by the largest value of the vector, and take the normalized activations as the assignment strengths, i.e.

$$P_j^i = A_j^i / max(A^i) \tag{2}$$

Where A^i means the activation values for the ith region, and A_j^i is the jth element of A^i.

Although the activations are sparse, there are still quite a lot of non-zero values inside. We only keep the assignment strengths with large values, and modify Eq.(2) as Eq.(3):

$$P_j^i = \begin{cases} 0 & \text{if } A_j^i < mean(A^i) \\ A_j^i / max(A^i) & \text{if } A_j^i \geq mean(A^i) \end{cases} \tag{3}$$

Motivated by the spatial pyramid matching (SPM) scheme [15], we further propose a spatial variant of BoSP, called S-BoSP. The specific procedure is: we divide the image equally into multiple sub-regions (9 regions in 3 rows and 3 columns in this paper), compute BoSP inside each sub-region and concentrate them into a single feature vector. Therefore, the dimension of S-BoSP is 9 times the one of BoSP. For simplicity, we conduct the partitioning process on feature maps, instead of the input images. As the size of feature maps for the pool$_5$ layer of VGG network is 7×7 , we need to divide the feature maps in an overlapping pattern, i.e. some of the surrogate parts would exist in multiple sub-regions.

3.2 Scale Pooling

The BoSP/S-BoSP described above only concern the spatial units at the finest level, and handle them in a disjoint way, which means to sample and assign regions in input images with fixed size and position. However, the objects may appear in different shapes, positions and scales, the independent processing of the spatial units may capture different parts of the same object and is inferior to assign the objects of different scales, thus makes the resulting feature less discriminative. To address this problem, we propose a scale pooling technique, which improves the assignment of objects with different scales and deformations by handling

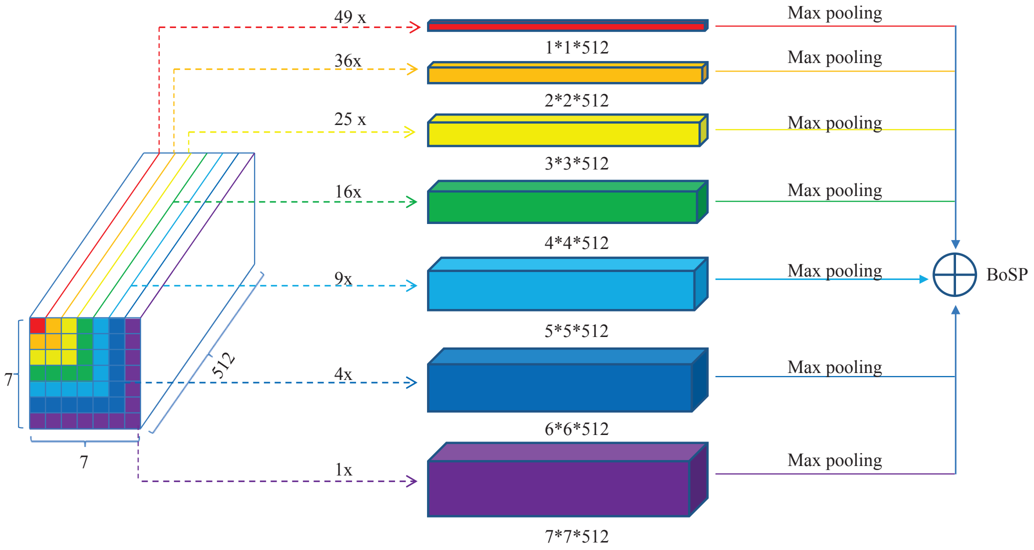

Figure 2: The illustration of scale pooling technique for BoSP (We can extract different number of features from 7 scales. For example, there are 49 red strips for the smallest scale, and only 1 purple strip for the largest scale. Then we apply max pooling on the features inside each scale and add them up to form the final feature).

regions of different sizes and positions, together with max pooling operations inside each region. The procedure of scale pooling is illustrated in Figure 2.

We divide the activations from the pool$_5$ layer into 7 scales for BoSP. Under different scales, we derive spatial units of different numbers and different sizes. For the sake of clarity, we name the derived spatial units as coarse spatial units, the coarse spatial units in scale 1 correspond to the fine spatial units. Under scale i ($i \in [1,7]$), we can derive $(8-i)^2$ coarse spatial units, each coarse unit contains i^2 fine spatial units, therefore, the total number of the coarse spatial units is $\sum_{i=1}^{7}(8-i)^2 = 140$. Next, we pool these coarse spatial units and calculate their assignment strengths for the surrogate parts utilizing Eq.(3). In this paper, we utilize max pooling operation inside each coarse spatial unit since it has been verified to be superior for capturing invariances in image-like data [25]. Finally, we sum the assignment strengths of the coarse spatial units under different scales together to form the refined assignment strengths for the image. For S-BoSP, we utilize the scale pooling schemes inside each sub-region, and then concentrate the features together.

The scale pooling scheme is proposed to handle the image regions of different sizes and different positions, making the assignment be more relaxable. Also, the scheme is robust to object deformation inside each coarse spatial unit. Furthermore, the introduction of scale pooling would not enlarge the feature dimension and does not affect the efficiency greatly.

3.3 Global Constrained Augmentation

Given an input image, we can extract its global feature after resizing it to 224×224 for VGG. However, in some cases one global feature would not be discriminative enough to classify images, and it is mostly beneficial to utilize data augmentation technique. Without extra data, one common approach of data augmentation is to generate numerous sub-images from the input image, and average the sub-image features as the augmented image feature. Although this approach could extract more information from one image, it only considers individual parts of the input image, and fails to handle the input image entirely. To make a more comprehensive prediction, we add a global constraint term upon the prediction of augmented features.

The specific procedure is: given an input image, we first resize it to 224×224, and extract the global feature. This feature focuses more on the entire image, and we can achieve the global prediction based on it, denoted as Pre_{global}; Next, we resize the image to make the smallest side equal S while keeping its ratio, and crop regions of 224×224 with stride of 32 pixels. Thereby, we formulate several sub-images from the input image, each sub-image may only contain part of the original object. The image feature is the average of the sub-image features, and this feature concerns more about parts of the image, and based on it, we make the part prediction, denoted as Pre_{part}. The final prediction is the product of the global prediction and part prediction:

$$Pre_{fusion} = Pre_{global} \times Pre_{part} \tag{4}$$

As our feature is derived from the convolutional layers, the input image could be of any size, and we do not need to explicitly crop sub-images. In practice, we only need to input the resized image once to extract the augmented BoSP/S-BoSP.

4 Experiments

To assess the performance of our method, we carry out a series of experiments on four datasets, Caltech101 [8], Oxford 102 Flowers (referred to as Oxford102) [19], MIT Indoor67 (referred to as Indoor67) [23] and SUN397 [29], which cover several popular topics in image classification, i.e. generic object classification, fine-grained object classification, and scene classification. The details of the datasets are described below:

Caltech101 has 102 classes (101 object categories and a background class) and a total of 9144 images, and the image number per category ranges from 31 to 800. We randomly select 30 images per class for training and test on up to 50 images per class. There are 44 ‘overlap’ images of the Caltech101 dataset and ImageNet training data. We exclude these images from the test set.

Oxford102 contains 8189 images for 102 flower categories, and each category contains 40 to 258 images under various scales, pose and illuminations. We use 20 images per class for training and the rest for testing.

Indoor67 consists of 67 indoor categories and a total of 15620 images. The standard split for this dataset consists of 80 training and 20 test images per category.

SUN397 is a large scale and general scene dataset, which contains more than 100K images for 397 scene categories. Each category has at least 100 images. The training and test splits are publicly available from [29], where each split has 50 training and 50 test images per category. The classification accuracy is reported by averaging the results of the 10 public splits.

In our scheme, VGG Net-D[27] is employed as the pre-trained CNN model to extract deep features. The model is implemented by Caffe [14] package. For simplicity, pre-trained model weights are kept fixed without fine-tuning. We employ the regularized logistic regression(LR) classifier for all the experiments owing to its high efficiency, and the LR classifier is implemented by utilizing the open source library: LIBLINEAR [7]. For fair comparisons, we set a fixed regularization term $\lambda = 20$ [1]. All of the BoSP/S-BoSP and CNN

[1] In preliminary test on Caltech101 involving λ, we found that the accuracy was robust to small changes (e.g. for the global BoSP feature, λ/accuracy: 18/91.68%, 20/91.65%, 22/91.69% and even 60/91.65%) and so 20 was chosen as a representative value.

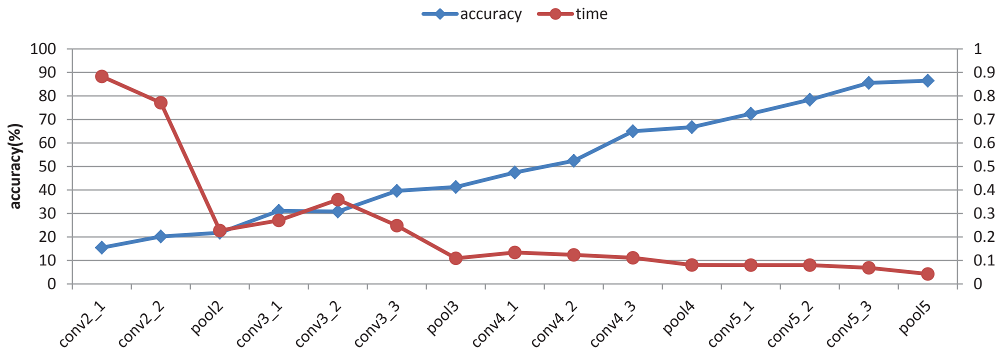

Figure 3: The performances of BoSP for each layer of the VGG network.

features are L2 normalized before sending into the classifier. For the data augmentation, we resize the images of Caltech101 to let its smallest side S equals 256, while $S = 512$ for Oxford102, Indoor67 and SUN397.

4.1 Analysis of our method

In the following, we first verify the advantages to derive BoSP from the pool$_5$ layer, and then demonstrate the effectiveness of our proposed scale pooling and global constrained augmentation schemes.

4.1.1 The comparison of BoSP from different layers

The proposed BoSP is derived from convolutional layers, and we can generate multiple BoSP features from different layers of the network. Intuitively, deeper layer activations contain more semantic information compared to shallower layer activations, thus could get better performance. To verify this, we evaluated the accuracy and efficiency of BoSP from different layers on Caltech101 dataset.

From Figure 3, we can observe that, in terms of accuracy, the performance of BoSP would generally increase with the layer depth, in which the feature derived from pool$_5$ layer achieves the best result. This phenomenon confirms our assumption for the advantage of deeper layers. As for the efficiency, it also increases with the layer depth, since deeper feature maps are smaller than shallower ones, therefore we need to assign fewer spatial units. The merits of high layers verify that, it is reasonable to extract our features from the pool$_5$ layer.

4.1.2 Evaluation of the Scale Pooling

In this part, we aim to evaluate the benefits brought by the proposed scale pooling scheme, and compare our proposed feature with commonly used CNN feature. All of the features in this part are extracted after resizing the images to 224×224.

The results in Table 1 highlight the advantage of the scale pooling scheme: it could improve the performance of BoSP/S-BoSP without enlarging the feature dimension, and the improvement can be very large. For example, scale pooling increases the BoSP features of Caltech101 and Oxford102 by more than 3%. However, the improvement brought by scale pooling varies with the feature. Generally, scale pooling is more effective for BoSP than S-BoSP. For instance, for Caltech101, scale pooling achieves more than 3% improvement for BoSP (from 88.28% to 91.65%), while only 0.4% increase for S-BoSP (from 93.59% to 93.99%). This is because, for BoSP, we can utilize scale pooling across the whole feature

Table 1: The comparison of BoSP/S-BoSP with/without scale pooling and the CNN feature extracted from the last fully-connected layer. The feature without scale pooling is marked with *

	Caltech101	Oxford102	Indoor67	SUN397	Dim
CNN	89.22%	80.60%	68.06%	53.26%	4096
*BoSP	88.28%	81.28%	69.47%	53.68%	512
BoSP	91.65%	**85.98%**	70.44%	54.12%	512
*S-BoSP	93.59%	83.88%	70.30%	54.52%	4608
S-BoSP	**93.99%**	85.54%	**71.19%**	**55.41%**	4608

Table 2: The comparison of the different predictions on CNN, BoSP and S-BoSP. Pre_{global}: prediction of the global feature; Pre_{part}: prediction of the augmented feature in the traditional way; Pre_{fusion}: prediction of the augmented feature in our proposed global constrained augmentation way.

	Caltech101		Oxford102		Indoor67		SUN397	
	BoSP	S-BoSP	BoSP	S-BoSP	BoSP	S-BoSP	BoSP	S-BoSP
Pre_{global}	91.65%	93.99%	85.98%	85.54%	70.44%	71.19%	54.12%	55.41%
Pre_{part}	91.68%	93.62%	92.14%	91.38%	74.78%	75.07%	58.91%	60.10%
Pre_{fusion}	92.75%	**94.52%**	**92.23%**	91.75%	77.69%	**77.91%**	61.58%	**62.95%**

map, and generate more coarse spatial units from seven scales. But for S-BoSP, as we have divided the feature map into different regions, we can only generate coarse spatial units from three scales.

On top of that, the results also demonstrate the effectiveness of the spatial scheme, since most of the S-BoSP performs better than the corresponding BoSP.

We further listed the comparison of the proposed BoSP/S-BoSP with commonly-used CNN feature, and found that, the proposed BoSP/S-BoSP from the pool_5 layer achieve considerably better performance than the activations from the fully-connected layer, indicating that the activations from convolutional layers also contain quite discriminative information for the classification task. Specifically, the BoSP for Oxford102 outperforms the corresponding CNN by more than 5%, and has a much lower dimension (e.g. 512D vs. 4096D).

4.1.3 Evaluation of the global constrained augmentation

Given an image, we can extract the global feature after resizing it to 224×224. However, in some cases, extracting only one feature from the input image is not sufficient to make a good prediction. Therefore, it is common to employ the data augmentation technique to enhance the performance. The traditional image augmentation scheme is to generate multiple sub-images from the original image, and average their features as the augmented image feature. This feature focuses more on 'parts' of the image. In this paper, we proposed a global constrained augmentation method to incorporate the predictions from both part-based feature and global-based feature, and evaluated this method on the proposed BoSP/S-BoSP features.

From the results in Table 2, we can first observe the effectiveness of the data augmentation, since most of the part-based predictions outperform the corresponding global-based predictions, in which the improvement of BoSP on Oxford102 can be more than 6%. This

Table 3: The comparison with the state-of-the-art

Method	Base Model	Dim	Caltech101	Oxford102	Indoor67	SUN397
Decaf [6]	AlexNet	4096	86.91%	-	-	40.94%
Places-CNN [35]	AlexNet	4096	65.18%	-	68.24%	54.32%
Hybrid-CNN [35]	AlexNet	4096	84.79%	-	70.80%	53.86%
MPP [33]	AlexNet	65536	-	91.28%	75.67%	-
MsML [22]	AlexNet	51456	-	88.39%	-	-
MsML+ [22]	AlexNet	134016	-	89.45%	-	-
SPP [12]	OverFeat	4096	93.42%	-	-	-
Deep Optimized [2]	OverFeat	4096	-	91.3%	71.3%	56%
VGG [27]	VGG-D&E	12288	92.7%	-	-	-
ONE [30]	VGG-E	4096	-	86.82%	70.13%	54.87%
CrossLayer [17]	VGG-E	262144	-	-	74.4%	-
Deep19-DAG [32]	VGG-E	9216	-	-	77.5%	56.2%
FV-CNN [4]	VGG-E	65536	-	-	**81%**	-
NML [26]	VGG-D	4096	-	84.3%	-	-
BoE [16]	VGG-D	5000	-	-	77.63%	-
BoSP (Pre$_{fusion}$)	VGG-D	512	92.75%	**92.23%**	77.69%	61.58%
S-BoSP (Pre$_{fusion}$)	VGG-D	4608	**94.52%**	91.75%	77.91%	**62.95%**

demonstrates the necessity of the data augmentation.

Furthermore, we can also confirm the advantage of our proposed global constrained augmentation scheme. Regardless of the differences of the global-based prediction and part-based prediction, it is always beneficial to incorporate them in our proposed method. Specifically, the improvement on the predictions of Indoor67 and SUN397 can be about 3%.

4.2 Comparison with the state-of-the-art

In Table 3, we compare our scheme with several published state-of-the-art methods which also utilized the CNN structure.

We can notice that, for generic object classification, our proposed method has the top accuracy on the Caltech101 dataset. The previous published result of VGG on Caltech101 dataset is 92.7%, which is achieved by concentrating the top-layer features from two models (i.e. VGG Net-D and VGG Net-E) under three scales ($S = 256, 384, 512$). In contrast, the proposed BoSP from the pool$_5$ layer of one model (VGG Net-D) in our scheme could get slightly better result with it (92.75% vs. 92.7%), while coming in lower dimension (512-D). The S-BoSP with global constrained augmentation further raise the state-of-the-art from 93.42% to 94.52%, which verifies the effectiveness of our scheme.

For fine-grained object classification, the spatial scheme does not help on the Oxford102 dataset, as S-BoSP delivers inferior performance compared to BoSP. This is possibly because the smaller parts of the fine-grained objects are more likely to be similar and do not distinguish well. Nevertheless, both BoSP and S-BoSP perform better than previous state-of-the-art result. Notably, BoSP achieves an improvement over the previous best result, from 91.3% to 92.23%, with the dimensionality of 512, demonstrating the high discriminatory power of our feature.

For indoor scene classification, our proposed method yields comparable results with the previous best performance on Indoor67 dataset, while having lower feature dimension. Currently, the FV-CNN employed in [4] achieves the state-of-the-art, with the accuracy of 81%. However, compared with our proposed BoSP/S-BoSP, the dimension of FV-CNN is much larger (65536 vs. 512/4608), which may be disadvantageous in large scale application.

For the large scale, general scene classification in the SUN397 dataset, our scheme achieves significantly better performance than the currently best result, improving the state-of-the-art from 56.2% to 62.95%.

5 Conclusion and Future Work

In this paper, we derive a new feature from convolutional layers, by densely sampling and assigning the regions in input images to the surrogate parts. The feature can be efficiently extracted and is highly discriminative. We also propose two schemes: scale pooling and global constrained augmentation, to further improve the performance. Through extensive experiments, we have shown that the proposed feature shows state-of-the-art classification performance at a low computational cost.

In the future, we will further exploit the convolutional activations in two possible directions: (1) we would explore the semantic meaning of the surrogate parts, and utilize the semantically meaningful surrogate parts to help the classification; (2) different layers contain different levels of detail, and we may be able to combine the BoSP/S-BoSP from different layers to further boost the performance.

References

[1] Pulkit Agrawal, Ross Girshick, and Jitendra Malik. Analyzing the performance of multilayer neural networks for object recognition. In *ECCV*, 2014.

[2] Hossein Azizpour, Ali Sharif Razavian, Josephine Sullivan, Atsuto Maki, and Stefan Carlsson. From generic to specific deep representations for visual recognition. In *CVPRW*, 2015.

[3] Artem Babenko and Victor Lempitsky. Aggregating deep convolutional features for image retrieval. In *ICCV*, 2015.

[4] Mircea Cimpoi, Subhransu Maji, and Andrea Vedaldi. Deep filter banks for texture recognition and segmentation. In *CVPR*, 2015.

[5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.

[6] Jeff Donahue, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng, and Trevor Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. In *ICML*, 2014.

[7] Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. Liblinear: A library for large linear classification. *JMLR*, 9:1871–1874, 2008.

[8] Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. *CVIU*, 106(1): 59–70, 2007.

[9] Yunchao Gong, Liwei Wang, Ruiqi Guo, and Svetlana Lazebnik. Multi-scale orderless pooling of deep convolutional activation features. In *ECCV*, 2014.

[10] Yanming Guo, Songyang Lao, Yu Liu, Liang Bai, Shi Liu, and Michael S Lew. Convolutional neural networks features: Principal pyramidal convolution. In *PCM*, 2015.

[11] Yanming Guo, Yu Liu, Ard Oerlemans, Songyang Lao, Song Wu, and Michael S Lew. Deep learning for visual understanding: A review. *Neurocomputing*, 187:27–48, 2016.

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. In *ECCV*, 2014.

[13] Hervé Jégou, Florent Perronnin, Matthijs Douze, Javier Sanchez, Pablo Perez, and Cordelia Schmid. Aggregating local image descriptors into compact codes. *PAMI*, 34(9):1704–1716, 2012.

[14] 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 MM*, 2014.

[15] Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR*, 2006.

[16] Yao Li, Lingqiao Liu, Chunhua Shen, and Anton van den Hengel. Mid-level deep pattern mining. In *CVPR*, 2015.

[17] Lingqiao Liu, Chunhua Shen, and Anton van den Hengel. The treasure beneath convolutional layers: Cross-convolutional-layer pooling for image classification. In *CVPR*, 2015.

[18] Joe Yue-Hei Ng, Fan Yang, and Larry S Davis. Exploiting local features from deep networks for image retrieval. In *CVPRW*, 2015.

[19] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *ICVGIP*, 2008.

[20] Maxime Oquab, Leon Bottou, Ivan Laptev, and Josef Sivic. Learning and transferring mid-level image representations using convolutional neural networks. In *CVPR*, 2014.

[21] Florent Perronnin and Christopher Dance. Fisher kernels on visual vocabularies for image categorization. In *CVPR*, 2007.

[22] Qi Qian, Rong Jin, Shenghuo Zhu, and Yuanqing Lin. Fine-grained visual categorization via multi-stage metric learning. In *CVPR*, 2015.

[23] Ariadna Quattoni and Antonio Torralba. Recognizing indoor scenes. In *CVPR*, 2009.

[24] Ali S Razavian, Hossein Azizpour, Josephine Sullivan, and Stefan Carlsson. Cnn features off-the-shelf: an astounding baseline for recognition. In *CVPR*, 2014.

[25] Dominik Scherer, Andreas Müller, and Sven Behnke. Evaluation of pooling operations in convolutional architectures for object recognition. In *Artificial Neural Networks–ICANN 2010*, 2010.

[26] Gaurav Sharma and Bernt Schiele. Scalable nonlinear embeddings for semantic category-based image retrieval. In *ICCV*, 2015.

[27] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015.

[28] Josef Sivic and Andrew Zisserman. Video google: A text retrieval approach to object matching in videos. In *ICCV*, 2003.

[29] Jianxiong Xiao, James Hays, Krista Ehinger, Aude Oliva, Antonio Torralba, et al. Sun database: Large-scale scene recognition from abbey to zoo. In *CVPR*, 2010.

[30] Lingxi Xie, Richang Hong, Bo Zhang, and Qi Tian. Image classification and retrieval are one. In *ICMR*, 2015.

[31] Wei Xiu-Shen, Gao Bin-Bin, and Wu Jianxin. Deep spatial pyramid ensemble for cultural event recognition. In *ICCV Workshop*, 2015.

[32] Songfan Yang and Deva Ramanan. Multi-scale recognition with dag-cnns. In *ICCV*, 2015.

[33] Donggeun Yoo, Sunggyun Park, Joon-Young Lee, and In Kweon. Multi-scale pyramid pooling for deep convolutional representation. In *CVPRW*, 2015.

[34] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *ECCV*, 2014.

[35] Bolei Zhou, Agata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. Learning deep features for scene recognition using places database. In *NIPS*, 2014.