Deep Spatial Pyramid Match Kernel for Scene Classification

Shikha Gupta[1], Deepak Kumar Pradhan[2], Dileep Aroor Dinesh[1] and Veena Thenkanidiyoor[2]
[1]*School of Computing and EE, Indian Institute of Technology, Mandi, H.P., India*
[2]*Department of CSE, National Institute of Technology Goa, Ponda, Goa, India*

Keywords: Scene Classification, Dynamic Kernel, Set of Varying Length Feature Map, Support Vector Machine, Convolutional Neural Network, Deep Spatial Pyramid Match Kernel.

Abstract: Several works have shown that Convolutional Neural Networks (CNNs) can be easily adapted to different datasets and tasks. However, for extracting the deep features from these pre-trained deep CNNs a fixed-size (e.g., 227×227) input image is mandatory. Now the state-of-the-art datasets like MIT-67 and SUN-397 come with images of different sizes. Usage of CNNs for these datasets enforces the user to bring different sized images to a fixed size either by reducing or enlarging the images. The curiosity is obvious that "Isn't the conversion to fixed size image is lossy ?". In this work, we provide a mechanism to keep these lossy fixed size images aloof and process the images in its original form to get set of varying size deep feature maps, hence being lossless. We also propose deep spatial pyramid match kernel (DSPMK) which amalgamates set of varying size deep feature maps and computes a matching score between the samples. Proposed DSPMK act as a dynamic kernel in the classification framework of scene dataset using support vector machine. We demonstrated the effectiveness of combining the power of varying size CNN-based set of deep feature maps with dynamic kernel by achieving state-of-the-art results for high-level visual recognition tasks such as scene classification on standard datasets like MIT67 and SUN397.

1 INTRODUCTION

CNNs have been popular these days for their applicability to wide range of tasks, such as object recognition (Simonyan and Zisserman, 2014), (Girshick et al., 2014), (Chatfield et al., 2014), image segmentation (Kang and Wang, 2014), image retrieval (Zhao et al., 2015), scene classification (He et al., 2015), (Yoo et al., 2014) and so on. Spectacular results for the state-of-the-art tasks are mainly because of powerful feature representation learnt from CNNs. Scene image classification, being the most basic and important aspect of computer vision, has received a high degree of attention among the researchers. An important issue in scene image classification is intra-class variability *i.e*, the images of a particular class differ so much in their visual appearance and inter-class similarity *i.e*, images of different class are very much confusable and composed by the similar concepts. For addressing these issues many deep CNNs such as AlexNet (Krizhevsky et al., 2012), GoogLeNet (Szegedy et al., 2015) and VGGNet-16 (Simonyan and Zisserman, 2014) have already been trained on datasets like Places-205, Places-365 (Zhou et al., 2017) and ImageNet (Deng et al., 2009) for image classification tasks. These deep CNNs can be adapted and retrained for other datasets and tasks with a little fine-tuning. In all such cases, features obtained from pre-trained or fine-tuned CNNs are used to build fully connected neural network or SVM-based classifier. These CNNs also became popular to greater extent as they are useful in providing base architecture and features for many other similar tasks than the one for which they are trained. For example, AlexNet (Krizhevsky et al., 2012) is trained for object recognition. However, (Mandar et al., 2015) used the features for scene classification by further enhancing through Fisher encoding. These CNNs require images to be input in a fixed size. For example the AlexNet accepts images of size "227×227". However the state-of-the-art datasets like SUN397 (Xiao et al., 2010) or MIT-67 indoor (Quattoni and Torralba, 2009) scene datasets comprise of varying sized images which are much larger than "227×227". The conventional approach to use these CNNs is to resize the arbitrary-sized images to a fixed size. This results in loss of information of the image before feeding it to the CNN for feature extraction. The performance of classification

Gupta, S., Pradhan, D., Dinesh, D. and Thenkanidiyoor, V.
Deep Spatial Pyramid Match Kernel for Scene Classification.
DOI: 10.5220/0006596101410148
In *Proceedings of the 7th International Conference on Pattern Recognition Applications and Methods (ICPRAM 2018)*, pages 141-148
ISBN: 978-989-758-276-9

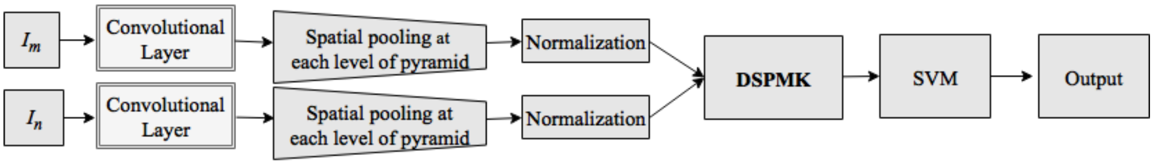

Figure 1: Block diagram of our proposed approach.

gets compromised due to such practice, which is evident from (He et al., 2015). To avoid any such information loss, many researchers explored different approaches to feed varying sized images (*i.e*, original size of the image) to CNN. (He et al., 2015) eliminates the requirement of fixed size image by including a spatial pyramid pooling (SPP) layer in CNN and named the network as SPP-Net. (Gao et al., 2015) follows the same approach by further processing the outputs of convolutional layer to encode it into a single vector using either vector of locally aggregated descriptor (VLAD) (Gong et al., 2014) or Fisher vector (FV) (Yoo et al., 2014), by building a Gaussian mixture model (GMM).

The research work presented in this paper focuses on giving the images in their original size as input to the CNN and then take sets of deep feature maps from the last convolutional layer. As we learn convolutional layers are the indispensable part of CNN and responsible for generating discriminative features. These deep feature maps are of varying size with respect to the corresponding original image size. They contain more spatial information compared to the activation of the fully connected layers, as fully connected layer destroy the spatial content present in the convolutional layer features. These sets of varying size deep feature maps are used to build support vector machine (SVM) based classifiers for varying length pattern classification. There are two approaches to varying length pattern classification using SVMs. In the first approach, a set of varying size deep feature maps is first mapped onto a fixed length pattern as in (Gao et al., 2015), and then a kernel for fixed length pattern is used to build the SVM-based classifier. In the second approach a suitable kernel for set of varying size deep feature maps is designed. Then it is used to build SVM-based classifier. The kernel designed for set of varying length feature vectors are called dynamic kernels (Dileep and Chandra Sekhar, 2014). The dynamic kernels in (Lazebnik et al., 2006), (Dileep and Chandra Sekhar, 2014), (Gupta et al., 2016a) and (Gupta et al., 2016b) show promising results for classification of varying size image and speech signals.

In this paper, we focus on dynamic kernel based SVMs for classification of set of varying size deep feature maps obtained from convolutional neural network. We propose deep spatial pyramid match kernel (DSPMK) for SVM-based classification of images represented as set of varying size deep feature maps inspired by the spatial pyramid match kernel of (Lazebnik et al., 2006). The entire process is shown in block diagram of Figure 1. In this block diagram, two images I_m and I_n of arbitrary size are passed as input to the convolutional layer of deep CNN to obtain set of deep feature maps. Size of feature maps obtain for image I_m can be different from size of feature maps obtain for I_n. These varying size deep feature maps for two images are further spatially divided and sum-pooled at each level of pyramid. These spatially pooled feature maps are normalized to obtain the probability vectors. Then, we propose to compute a matching score between probability vector representation of I_m and I_n using DSPMK. A DSPMK-based SVM classifier is used to learn the association of sets of deep feature maps with the class label. The main contribution of this paper toward exploring set of varying size deep feature maps obtained from deep CNN is in computing DSPMK for building SVM-based classifier. This is in contrast to (Gao et al., 2015) where varying size deep feature maps are encoded to fixed length Fisher vector and then linear kernel-based SVM is used for classification. Salient feature of the proposed DSPMK is that it works for different sized images by building spatial pyramid of $L+1$ levels ranging from 0, 1 to L using set of varying size deep feature maps.

This paper is organized as follows: In Section 2, a review of related approach for scene image classification using CNN-based features is presented. The proposed DSPMK for set of varying size deep feature map is described in Section 3. In Section 4 the experimental studies using the proposed approach on scene classification is presented. The conclusion is presented in Section 5.

2 RELATED WORK

In this section, we review the state-of-the-art approaches for classification of varying length feature vector representation. In the last decade, traditional hand engineered low-level image descriptors like the histogram of oriented gradient (HOG) (Dalal and Triggs, 2005) and scale invariant feature transform (SIFT) (Lowe, 2004) were popular. These descriptors result in the set of local feature vector representation for images. An SVM-based classifier for such representation can be built using the standard kernels such as Gaussian kernel by mapping local feature vectors onto a fixed dimensional representation. One of the commonly used fixed-dimensional vector representations of an image is bag-of-visual words (BOVW) representation (Lazebnik et al., 2006). The BOVW representation is a fixed-dimensional vector with frequencies of occurrence of visual words in an image as its elements. A limitation of the BOVW representation is that there is the loss of spatial information. Alternatively, SVM-based classifiers can be built for scene classification using the dynamic kernels that are designed for the varying size sets of local feature vectors (Lazebnik et al., 2006) (Thenkanidiyoor et al., 2017). One of the approaches to the design of dynamic kernels is the matching based approach that involves computing value of kernel function between two sets of local feature vectors by matching the local feature vectors. Spatial pyramid match kernel (Lazebnik et al., 2006), GMM-based intermediate matching kernel (Dileep and Chandra Sekhar, 2014) and segment-level pyramid match kernel (Gupta et al., 2016a) are some of the matching based dynamic kernels for classification of varying size image and speech signals. With the advancement of deep convolutional neural networks, traditional features and related techniques are being replaced by CNN-based features with linear kernel based SVM classifier. Due to the strong feature capturing ability of CNN trained on large datasets, one can directly use features from fully connected layers of deep CNNs to build SVM-based classifier and achieve better performance than traditional methods (Zhou et al., 2014). Some researchers also tried to encode CNN-based features into a new representation. (Mandar et al., 2015) have encoded the features from fully connected layer of deep CNN into a Bag-of-Semantics (BoS). This BoS is then summarized in "semantic Fisher vector" representation.

For extracting fully connected features from pre-trained deep CNNs, the input images are resized to a fixed size. The coercive nature of fully-connected layer expects a fixed length representation of input, whereas the convolution process is not constrained with fixed length representation. In other words, the necessity of fixed size image as input to deep CNNs is an indirect requirement of the fully-connected layer. The process of resizing the images to a fixed size lead to a loss in information (He et al., 2015). On the other hand, any arbitrary sized input image can be fed into the convolutional layers of CNN which results in arbitrary sized deep feature maps. These deep feature maps correspond to the strongest responses from filters of convolution layer and at the same time they preserves spatial information. The similar idea can be observed from the work in (Yoo et al., 2015) and (He et al., 2015). These works have considered image scaled pyramid and spatial pyramid approaches to incorporate the concept information of images into the feature maps. (Yoo et al., 2015) have focused in considering scale characteristics over activation maps. Dense activation maps are obtained for a image in pyramid of seven layers. Each layer in this case consists of activation map from differently scaled image. These multi scale dense activation maps are aggregated in a Fisher kernel framework. (He et al., 2015) have followed spatial pyramid approach to eliminate the requirement of fixed size images input in CNNs. Here, the CNN is fed with original image size. However, in (Yoo et al., 2015) the CNN is fed with differently scaled images. (Gao et al., 2015) also followed the same approach and fed the arbitrary sized image to CNN. However, the approach for spatial pyramid is different. Here, a GMM using fixed size vector representation obtained from spatial pyramid pooling is built to generate Fisher vectors. Finally, all the Fishers vectors are concatenated to form a fixed dimensional representation. Our work focuses on combining the power of dynamic kernel with CNN-based set of varying size deep feature maps to obtain a matching score between a pair of images of different size. We propose to compute DSPMK instead of building GMM based dictionary. In the next Section, the proposed DSPMK for the set of varying size deep feature maps is presented.

3 DEEP SPATIAL PYRAMID MATCH KERNEL

In designing deep spatial pyramid match kernel (DSPMK), an image represented by a set of deep feature maps is decomposed into the pyramid of increasingly finer spatial regions. Here the size of deep feature maps obtains from last layer convolutional filters for the particular image is same but vary from one image to other as the images are fed to CNN in its original size. DSPMK between a pair of images is computed by matching the corresponding sum pooled feature vectors from spatially partitioned deep feature maps at each level of the pyramid. Let $\mathcal{D} = \{I_1, I_2, \ldots, I_m, \ldots, I_N\}$ be the set of all images in the dataset and'f' be the number of filters in last convo-

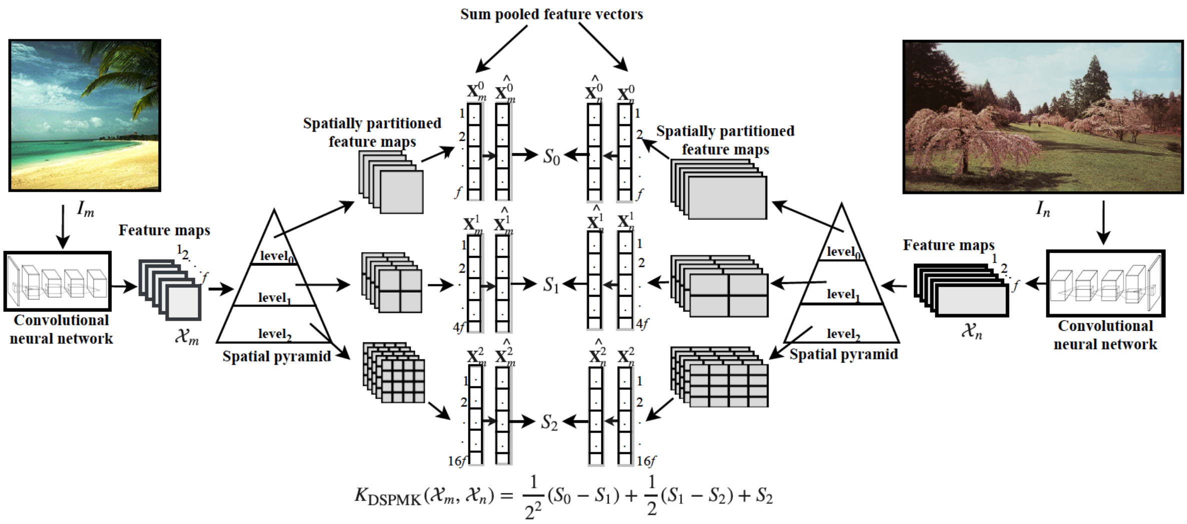
Sum pooled feature vectors
Spatially partitioned feature maps
Spatially partitioned feature maps
I_m
I_n
Feature maps
Feature maps
Convolutional neural network
Convolutional neural network
$\mathcal{X}_m$
$\mathcal{X}_n$
level$_0$
level$_1$
level$_2$
Spatial pyramid
Spatial pyramid
S_0
S_1
S_2
f
$4f$
$16f$
$K_{\text{DSPMK}}(\mathcal{X}_m, \mathcal{X}_n) = \frac{1}{2^2}(S_0 - S_1) + \frac{1}{2}(S_1 - S_2) + S_2$

Figure 2: Illustration of computing deep spatial pyramid match kernel between two different sized images I_m and I_n. Here, set of varying size deep feature maps $\mathcal{X}_m$ and $\mathcal{X}_n$ are computed from a pre-trained CNN. $\mathbf{X}_m^l$ and $\mathbf{X}_n^l$ are sum pooled deep feature representation of I_m and I_n which are computed from spatially partitioned deep feature maps at level l, (l=0 to 2). $\widehat{\mathbf{X}}_m^l$ and $\widehat{\mathbf{X}}_n^l$ are the probability vector representation of $\mathbf{X}_m^l$ and $\mathbf{X}_n^l$ obtained using (1). The intermediate similarity score at each level l ($i.e$, S_0, S_1 and S_2) is computed using equation (2).

lutional layer of pretrained deep CNN. Consider the mapping $\mathcal{F}$, which take input as original image and transform it to set of deep feature maps using convolutional layers of CNN. Mapping $\mathcal{F}$ is given as, $\mathcal{X}_m = \mathcal{F}(I_m)$.

At first phase, we feed the images to CNN in its original size. The size of images are different from each other in database and so as their size of deep feature maps. For image I_m, we have a set $\mathcal{X}_m$= $\{\mathbf{x}_{m1}, \mathbf{x}_{m2}, \mathbf{x}_{m3}, \ldots, \mathbf{x}_{mf}\}$ consisting of 'f' deep feature maps from mapping $\mathcal{F}$, where $\mathbf{x}_{mi} \in \mathbb{R}^{p_m \times q_m}$ and $p_m \times q_m$ is the size of each feature map obtained from last convolutional layer which varies according to the input image size. This leads to varying size deep feature map as shown in Figure 2 for images I_m and I_n.

In the second phase, a deep feature map is spatially partitioned into sub-regions forming deep spatial pyramid. At level-0 of pyramid, there will not be any spatial partition. At level-1, a deep feature map splits into 4 spatial regions corresponding to 4 quadrant, as shown in Figure 2. Let $L+1$ be the number of levels in spatial pyramid ranging from 0, 1 to L. At any level-l, a deep feature map $\mathbf{x}_{mi}$ is spatially split into 2^{2l} regions. The activation values of all the cells in a spatial region are sum pooled. At any level-l, activation values of cells in every spatial region of all the f deep feature maps are sum pooled and concatenated to form a vector $\mathbf{X}_m^l$ of size $f2^{2l} \times 1$ $i.e$, $\mathbf{X}_m^l$= $[X_{m1}^l, X_{m2}^l, \ldots, X_{mj}^l \ldots X_{m(f2^{2l})}]^\top$. This process is illustrated in Figure 2 by considering three levels, l = 0, 1 and 2.

In our proposed approach we considered three such levels. At level-0, ($i.e$, $l = 0$) the entire feature maps are sum pooled and we get $f \times 1$ dimensional vector representation. At level-1 ($i.e$, $l = 1$), the same feature maps are divided into four equal spatial regions. Each sum pooled spatial region of f feature maps results into a vector of $4f \times 1$ dimensional vector. Similarly, at level-2 ($i.e.$ $l = 2$), again the same feature maps are divided into sixteen equal spatial regions resulting into a vector of $16f \times 1$ dimensional vector. The $\mathbf{X}_m^l$ can now be seen as representation of image I_m at level-l of pyramid. At this stage, we propose to compute deep spatial pyramid match kernel (DSPMK) to match two images rather than deriving Fisher vector (FV) representation as in (Gao et al., 2015). Our proposed approach avoids building GMM to obtain FV and hence reduces the computation complexity as compared to (Gao et al., 2015). The process of computing DSPMK is motivated from spatial pyramid match kernel (SPMK) (Lazebnik et al., 2006). SPMK involves the histogram intersection function that match the frequency based image representation of two images at every levels of pyramid (Lazebnik et al., 2006). However, $\mathbf{X}_m^l$ is not the probability vector representation of image I_m. We sum normalize $\mathbf{X}_m^l$ to transform it into probability vector representation. Let $\mathbf{X}_m^l$ and $\mathbf{X}_n^l$ be the deep representation at level-l of two images I_m and I_n respectively. The probability vector representation of $\mathbf{X}_m^l$ and $\mathbf{X}_n^l$ is obtained as:

$$\widehat{\mathbf{X}}_m^l = \frac{\mathbf{X}_m^l}{\sum\limits_{j=1}^{f2^{2l}} X_{mj}^l} \ ,\widehat{\mathbf{X}}_n^l = \frac{\mathbf{X}_n^l}{\sum\limits_{j=1}^{f2^{2l}} X_{nj}^l} \qquad (1)$$

A histogram intersection function is used to compute intermediate matching score S_l between $\widehat{\mathbf{X}}_m^l$ and $\widehat{\mathbf{X}}_n^l$ at each level l as,

$$S_l = \sum_{j=1}^{f2^{2l}} min(\widehat{X}_{mj}^l, \widehat{X}_{nj}^l) \qquad (2)$$

Here, the matching score S_l found at level l also includes all the matches found at the finer level l + 1. Therefore, the number of new matches found at level l is given by $S_l - S_{l+1}$ for $l = 0, \dots, L$-1. The DSPMK is computed as a weighted sum of the number of new matches at different levels of the spatial pyramid. The weight associated with level l is set to $\frac{1}{2^{(L-l)}}$, which is inversely proportional to width of spatial regions at that level. The DSPMK kernel is computed as,

$$K_{\mathrm{DSPMK}}(\mathcal{X}_m, \mathcal{X}_n) = \sum_{l=0}^{L-1} \frac{1}{2^{L-l}}(S_l - S_{l+1}) + S_L \qquad (3)$$

The main advantages of using DSPMK is it works for different sized images. Secondly, it empowers the deep CNN-based set of varying size deep feature maps with dynamic kernel based SVM.

4 EXPERIMENTAL STUDIES

In this section, the effectiveness of the proposed dynamic kernel *i.e*, DSPMK is studied for scene classification task using SVM-based classifiers.

4.1 Datasets

We tested our proposed approach on four widely used scene classification datasets: MIT-8 Scene (Oliva and Torralba, 2001), Vogel-Schiele (VS) (Vogel and Schiele, 2004), MIT-67 (Quattoni and Torralba, 2009) and SUN-397 (Xiao et al., 2010).

MIT-8-scene dataset comprises of 2688 scene images belonging to 8 semantic classes, namely, ‘coast’, ‘forest’, ‘mountain’, ‘open-country’, ‘highway’, ‘inside-city’, ‘tall building’ and ‘street’. We follow the procedure of (Oliva and Torralba, 2001) and randomly select 100 images per class for training in each of the 5 trials and test on remaining images in each trial. The results presented correspond to the average classification accuracy for 5 trials.

Vogel-Schiele dataset comprises of 700 scene images of 6 semantic classes, namely, ‘coasts’, ‘forests’, ‘mountains’, ‘open-country’, ‘river’ and ‘sky-clouds’. The results presented for the studies using this dataset correspond to the average classification accuracy for stratified 5-fold cross validation.

MIT-67 indoor dataset comprises of 15,620 images with 67 indoor scene categories. It is a challenging dataset, because interclass variation between different classes is very less. The standard split (Quattoni and Torralba, 2009) for this dataset consist of approximately 80 training and 20 test example per class.

SUN397 is the large dataset for scene recognition. It includes 397 categories of indoor, urban and nature where each category has at least 100 images. The train and test splits are fixed and publicly available from (Xiao et al., 2010), where each split has 50 training and 50 testing images per category. We select the first five splits and the result presented is the average classification accuracy for 5 splits.

4.2 Experiment Details

In our studies, we have used three different CNN architectures namely AlexNet (Krizhevsky et al., 2012), GoogLeNet (Szegedy et al., 2015) and VGGNet-16 (Simonyan and Zisserman, 2014) for deep feature extraction which are pre-trained networks on ImageNet (Deng et al., 2009), Places205 and Places365 (Zhou et al., 2017) datasets. Reason for using the different network in our study is that ImageNet dataset is having mainly object centric images so it gives activations for object-like structures in the image, whereas Places205 and Places365 datasets comprises mostly scene images and CNNs trained on scene images are selective for landscapes, natural structure of scene with more spatial feature. In all these pre-trained CNN models weights are kept fixed without fine-tuning. These CNNs are employed without its fully-connected layers in our experiments so that input images of arbitrary size can be accepted. As discussed in Section 3, we have passed the original image of arbitrary size as input to deep CNNs and extracted set of varying size deep feature maps from last convolutional layer. The size of feature map depends on the number of filters f in last convolutional layer of deep CNN architecture and image size. The number of filters f, in last convolution layer of AlexNet, GoogLeNet and VGGNet-16 are 256, 1024 and 512 respectively. The architecture of these CNNs also differs from each other. So, feature map size will vary from image to image and architecture to architecture. DSPMK between varying size deep feature map for pair of images is computed as in Figure 2 using equation (1) to (3). We consider $L+1 = 3$ as the

Table 1: Comparison of classification accuracy (CA) (in %) with 95% confidence interval for the SVM based classifier using DSPMK on different datasets. Base features for the proposed approach are extracted from AlexNet (Krizhevsky et al., 2012), GoogLeNet (Szegedy et al., 2015) and VGGNet-16 (Simonyan and Zisserman, 2014) which are pre-trained deep network on ImageNet, Places-205 and Places-365 dataset. The highest accuracy of each column is marked in bold.

Different pre-trained deep CNN architectures used to build DSPMK-based SVM	MIT-8 scene	Vogel-Schiele	MIT-67	SUN-397
ImageNet-AlexNet (Krizhevsky et al., 2012)	93.52±0.13	79.46±0.23	62.46	45.46±0.12
Places205-AlexNet (Zhou et al., 2014)	93.56±0.12	82.21±0.25	62.24	53.21±0.23
Places365-AlexNet (Zhou et al., 2017)	94.15±0.11	82.90±0.31	66.67	55.43±0.24
ImageNet-GoogLeNet (Szegedy et al., 2015)	92.02±0.06	82.30±0.25	71.78	50.32±0.31
Places205-GoogLeNet (Zhou et al., 2014)	92.15±0.18	85.84±0.36	75.97	57.43±0.26
Places365-GoogLeNet (Zhou et al., 2017)	93.70±0.16	85.54±0.21	75.60	59.89±0.21
ImageNet-VGG (Simonyan and Zisserman, 2014)	93.90±0.07	84.62±0.31	75.78	53.67±0.25
Places205-VGG (Zhou et al., 2014)	94.54±0.03	**86.92±0.26**	**81.87**	61.86±0.24
Places365-VGG (Zhou et al., 2017)	**95.09±0.14**	84.68±0.28	77.76	**62.31**±0.25

Table 2: Comparison of classification accuracy (CA) (in %) with 95% confidence interval of proposed approach with state-of-the-art approaches on MIT-8 scene, Vogel-Schiele, MIT-67 Indoor and SUN-397 dataset. (SIFT: Scale invariant feature transform, IFK: Improved Fisher kernel, BoP: Bag of part, MOP: Multi-scale orderless pooling, FV: Fisher vector, DSP: Deep spatial pyramid, MPP: Multi-scale pyramid pooling, DSFL: Discriminative and shareable feature learning). The highest accuracy of each column is marked in bold.

Methods	Description	MIT-8-Scene	Vogel-Schiele	MIT-67	SUN-397
(Lowe, 2004)	SIFT+BOVW	79.13±0.13	67.49±0.21	45.86	24.82±0.34
(Juneja et al., 2013)	IFK+BoP	85.76±0.12	73.23±0.23	63.18	-
(Gong et al., 2014)	MOP-CNN	89.45±0.11	76.81±0.27	68.88	51.98±0.24
(Zhou et al., 2014)	Places-CNN-fc7	88.30±0.09	76.02±0.31	68.24	54.32±0.14
(Zhou et al., 2014)	Hybrid-CNN-fc7	91.23±0.04	78.56 ±0.21	70.80	53.86±0.21
(Mandar et al., 2015)	fc8-FV	88.43±0.08	79.56±0.23	72.86	54.40±0.30
(Gao et al., 2015)	VGGNet-16 + DSP	92.34±0.12	81.34±0.27	76.34	57.27±0.34
(Yoo et al., 2015)	MPP(Alex-fc7)+DSFL	-	-	80.78	-
Ours	DSPMK + VGGNet	**95.09±0.14**	**86.92±0.26**	**81.87**	**62.31±0.25**

number of levels in spatial pyramid. We consider LIBSVM (Chang and Lin, 2011) tool to build the DSPMK-based SVM classifier. Specifically, we uses one-against-the-rest approach for multi-class scene image classification. In SVM for building the classifier, we uses default value of trade-off parameter $C = 1$.

4.3 Results on Scene Classification

In this section, we present the experimental studies of scene image classification using proposed DSPMK-based SVM classifier and compare with state-of-the-art approaches. The scene classification accuracies for DSPMK-based SVMs are given in Table 1. Table 1 compares the classification accuracies of SVM using DSPMK which is constructed from varying size deep feature maps obtained from different pre-trained CNN models. It is seen that performance of SVM-based classifier with DSPMK obtained using deep features from VGGNet-16 is significantly better than that of SVM with DSPMK obtained using deep features from GoogLeNet and AlexNet. Reason being VGGNet-16 has very deep network compare to other architectures and it learns the hierarchical representation of visual data more efficiently.

Table 2 presents the comparison of scene image classification accuracy of proposed DSPMK-based SVM classifier with that of state-of-the-art approaches. From Table 2, it is seen that our proposed approach is giving better performance in comparison with traditional feature based approaches in (Lowe, 2004), (Juneja et al., 2013) and also with CNN-based approaches in (Zhou et al., 2014), (Gong et al., 2014), (Mandar et al., 2015), (Gao et al., 2015), (Yoo et al., 2015).

(Lowe, 2004) uses scale invariant feature transform (SIFT) descriptors for scene images feature representation. In (Lowe, 2004) paper, scene images are represented as set of local feature vectors, which are further converted into bag-of-visual word (BOVW) representation for classification using linear kernel based SVM classifier. (Juneja et al., 2013) uses the learned bag-of-part (BoP) representation and combine with improved Fisher vector for building SVM based classifier using linear kernel. (Gong et al., 2014) extracted CNN-based features from multiple scale of image at different levels and performs orderless vectors of locally aggregated descriptors (VLAD) pooling (Jégou et al., 2010) at every scale separately. The representations from different level are then concatenated to form a new representation known as multi-scale orderless pooling (MOP) which is used for training linear kernel based SVM classifier. (Zhou et al., 2014) uses more direct approach, where a large scale image dataset (Places dataset) is used for training the AlexNet architecture and extracted fully connected (fc7) layer feature from the trained network. The basic architecture of their Places-CNN is same as that of the AlexNet (Krizhevsky et al., 2012) trained on ImageNet. (Zhou et al., 2014) also trained a Hybrid-CNN, by combining the training data of Places dataset with ImageNet dataset. Here, features from fully connected (fc7) layer are then used for training linear SVM based classifier . (Mandar et al., 2015) obtained the semantic Fisher vector (FV) using standard Gaussian mixture encoding for CNN-based feature. Further linear kernel based SVM classifier is build using semantic FV for classification of scene images. (Gao et al., 2015) used the generative model based approach to build a dictionary on top of CNN feature maps. A FV representation for different spatial region of activation map is then obtained from the dictionary. A power and l_2 normalization is applied on the combined FV from different spatial region. A linear kernel based SVM classifier is then used for scene classification. (Yoo et al., 2015) combined the features from fc7 layer of AlexNet (Alex-fc7) and their complementary features named discriminative and shareable feature learning (DSFL). DSFL learns discriminative and shareable filters with a target dataset. The final image representation is used with the linear kernel based SVM classifier for the scene classification task.

In contrast to all the above briefly explained approaches, our proposed approach uses the image of arbitrary size and get the deep feature map of varying size without any loss of information. Secondly, we have proposed the deep spatial pyramid match kernel which handles the set of varying size deep feature maps and intend to incorporate the local spatial information at the time of computing level wise matching score. Specifically, our proposed approach is very simple and discriminative in nature which outperforms the other CNN-based approach without combining any complementary features as in (Yoo et al., 2015). Our proposed approach reveals that for scene classification with complicated standard dataset good classification accuracy is achievable by using last convolutional layer features and DSPMK based-SVM. Proposed approach is independent of fully connected layer feature, capture the original size image, simple, memory efficient and take very less computation time in compare to state-of-the-art-approaches.

5 CONCLUSION

In this paper, we proposed DSPMK for enhancing the base features from last convolutional layer of CNN. DSPMK-based SVM classifies varying size scene images which are represented as the set of varying size deep feature maps. It is certain that better the feature the higher the performance. Our framework is equipped with a dynamic kernel which computes layer wise intermediate matching score and strengthens the matching procedure of convolutional layer features. The training of DSPMK-based SVM classifier consumes much lesser time than that of the training of GMM in (Gao et al., 2015). We have shown how the concepts in an image can be matched at the coarser level of the pyramid that is not matched in finer level. In our study, we have considered last convolutional layer features instead of fully connected layer features because fully connected layer restricts these features to fixed size and requires heavy computation time as it contains roughly 90% of the total parameter of CNN. Hence, convolutional layer features proved to be beneficial for us as images size of SUN-397 and MIT-67 datasets are also varying and large. The resized images will result in huge amount of concept and spatial information loss. Thus, convolutional filters will not be capable of capturing those resized concepts present in image. In terms of performance, our proposed approach achieves state-of-the-art results for standard scene classification datasets. In future, for capturing variations of the activations caused by the different size of concepts in an image and multi-scale deep spatial pyramid match kernel can be explored.

REFERENCES

Chang, C.-C. and Lin, C.-J. (2011). LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27.

Chatfield, K., Simonyan, K., Vedaldi, A., and Zisserman, A. (2014). Return of the devil in the details: Delving deep into convolutional nets. *arXiv preprint arXiv:1405.3531*.

Dalal, N. and Triggs, B. (2005). Histograms of oriented gradients for human detection. In *Computer Vision and*

Pattern Recognition (CVPR), 2005 IEEE Conference on, volume 1, pages 886–893.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 248–255. IEEE.

Dileep, A. D. and Chandra Sekhar, C. (2014). GMM-based intermediate matching kernel for classification of varying length patterns of long duration speech using support vector machines. *IEEE Transactions on Neural Networks and Learning Systems*, 25(8):1421–1432.

Gao, B.-B., Wei, X.-S., Wu, J., and Lin, W. (2015). Deep spatial pyramid: The devil is once again in the details. *CoRR*, abs/1504.05277.

Girshick, R., Donahue, J., Darrell, T., and Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587.

Gong, Y., Wang, L., Guo, R., and Lazebnik, S. (2014). Multi-scale orderless pooling of deep convolutional activation features. In *European conference on computer vision*, pages 392–407. Springer.

Gupta, S., Dileep, A. D., and Thenkanidiyoor, V. (2016a). Segment-level pyramid match kernels for the classification of varying length patterns of speech using svms. In *Signal Processing Conference (EUSIPCO), 2016 24th European*, pages 2030–2034. IEEE.

Gupta, S., Thenkanidiyoor, V., and Dinesh, D. A. (2016b). Segment-level probabilistic sequence kernel based support vector machines for classification of varying length patterns of speech. In *International Conference on Neural Information Processing*, pages 321–328. Springer.

He, K., Zhang, X., Ren, S., and Sun, J. (2015). Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 37(9):1904–1916.

Jégou, H., Douze, M., Schmid, C., and Pérez, P. (2010). Aggregating local descriptors into a compact image representation. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 3304–3311. IEEE.

Juneja, M., Vedaldi, A., Jawahar, C., and Zisserman, A. (2013). Blocks that shout: Distinctive parts for scene classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 923–930.

Kang, K. and Wang, X. (2014). Fully convolutional neural networks for crowd segmentation. *arXiv preprint arXiv:1411.4464*.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105.

Lazebnik, S., Schmid, C., and Ponce, J. (2006). Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Computer Vision and Pattern Recognition (CVPR), 2006 IEEE Conference on*, volume 2, pages 2169–2178.

Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110.

Mandar, Dixit, S., Chen, D., Gao, N., Rasiwasia, Nuno, and Vasconcelos (2015). Scene classification with semantic fisher vectors. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2974–2983.

Oliva, A. and Torralba, A. (2001). Modeling the shape of the scene: A holistic representation of the spatial envelope. *International journal of computer vision*, 42(3):145–175.

Quattoni, A. and Torralba, A. (2009). Recognizing indoor scenes. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 413–420. IEEE.

Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. (2015). Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9.

Thenkanidiyoor, V., Dileep, A. D., and Chandra Sekhar, C. (2017). Dynamic kernels based approaches to analysis of varying length patterns in speech and image processing tasks. In Amita Pal, S. K. P., editor, *Pattern Recognition and Big Data*. World Scientific.

Vogel, J. and Schiele, B. (2004). Natural scene retrieval based on a semantic modeling step. In *International Conference on Image and Video Retrieval*, pages 207–215. Springer.

Xiao, J., Hays, J., Ehinger, K. A., Oliva, A., and Torralba, A. (2010). Sun database: Large-scale scene recognition from abbey to zoo. In *Computer vision and pattern recognition (CVPR), 2010 IEEE conference on*, pages 3485–3492. IEEE.

Yoo, D., Park, S., Lee, J.-Y., and Kweon, I. S. (2014). Fisher kernel for deep neural activations. *arXiv preprint arXiv:1412.1628*.

Yoo, D., Park, S., Lee, J.-Y., and So Kweon, I. (2015). Multi-scale pyramid pooling for deep convolutional representation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 71–80.

Zhao, F., Huang, Y., Wang, L., and Tan, T. (2015). Deep semantic ranking based hashing for multi-label image retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1556–1564.

Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., and Torralba, A. (2017). Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Zhou, B., Lapedriza, A., Xiao, J., Torralba, A., and Oliva, A. (2014). Learning deep features for scene recognition using places database. In *Advances in neural information processing systems*, pages 487–495.