SPATIAL ENSEMBLE KERNEL LEARNING FOR SCENE CLASSIFICATION

Lei Zhang[1], *Xiantong Zhen*[2], *Qiujing Zhang*[1]

[1] College of Computer and Electronic Information, Guangdong University of Petrochemical Technology, Maoming, China

[2] School of Electronic and Information Engineering, Beihang University, Beijing,China

ABSTRACT

Scene recognition is one of the most important tasks in computer vision. Apart from appearance, spatial layout carries the crucial cue for discriminative representation. In this paper, we propose spatial ensemble kernel (SEK) learning, which enables fusion of multi-scale spatial information to achieve compact while discriminative representation of scenes. Based on the spatial pyramid, SEK combines the CNN features in each level of the pyramid in an ensemble and fuse them by kernels. By kernel approximation, we achieve Fourier feature embedding of CNN features in each scale, which establishes a nonlinear layer of the neural network with a cosine activation function. The parameters of the nonlinear layer can be learned jointly in one single optimization framework by supervised learning, which enables compact and discriminative feature representations. We show the effectiveness of the proposed SEK on two recent scene benchmark datasets, i.e., MIT indoor and SUN 397. The propose SEK produces high performance on two datasets which are competitive to state-of-the-art algorithms.

Index Terms— Spatial Ensemble Kernel, CNNs, Fourier Feature Embedding, Spatial Pyramid Kernel, Scene Classification

1. INTRODUCTION

Scene classification has significant meaning for many applications as autonomous driving [1], augmented reality [2], or geo-localizing archival imagery [3]. While a great amount of research has been conducted on robust scene representation, scene classification still remains extremely challenging for the varying appearance of the same place and similar looking among different scene environments.

Most approaches for scene classification adopt the similar framework as other tasks, such as representing scene images as orderless collections, which are aggregated into a single vector representation for the entire image such as bag-of-visual words [4], VLAD [5] or Fisher vector [6]. In particular, SIFT plus Fisher vector encoding has obtained the best ingredients for PASCAL VOC challenges 2012 [7].

Thanks to National Science Foundation of China (61571147), National Science Foundation of Heilongjiang (F2015027)

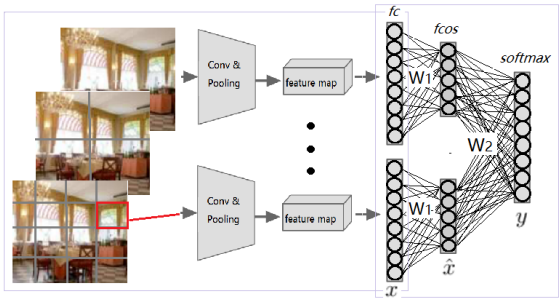

Fig. 1. Whole structure of CNNs & Spatial pyramid & cosine activation function

Due to GPU-based computation power together with large labelled image datasets, convolutional neural networks (CNNs) have emerged as powerful image representations for various category-level recognition tasks. [8] has demonstrated impressive performance on large scale object recognition. Thanks to the learned model as AlexNet, VGG and GoogLeNet, many methods directly extract descriptors from fully connected layer of these models and achieve soundable performance. Recently, Instead of fully connected layer feature, the convolutional features together with VLAD or Fisher vector are popular since the convolutional features are substantially less committed to a specific dataset than the fully connected layers. [9] combines VLAD into CNNs by adding a generalized VLAD layer. [10] realizes Fisher vector over *conv*5 and a collection of *fc*7 activations extracted from local crops or patches as semantic Fisher vector embedding. [11] extracts *fc*7 activations of local patches at multiple scale levels and names it as MOP-CNN.

In fact, scene classification is different from object recognition in that, iconic images of objects do not contain the richness and diversity of visual information that pictures of scenes and environments provide. It is the big obstruction when directly applying learned model from object classification task in scene classification.

Recently, transfer learning for scene classification is popular which can inherit the fine-tuned parameters from ImageNet for object recognition. Meanwhile, they refine the parameters in several layers to adapt to special scene classifi-

cation task. In addition, special scene-centric CNNs are established from millions of labeled scene images according to a standard CNN architecture in [12]. [12] shows deep learned representation of hierarchical organization from the dense and rich variety of natural scene image, which demonstrates that an object-centric network (using ImageNet) and a scene-centric network (using Places) learn different features.

Neither hand-crafted descriptor nor deep learned features from CNNs consider about the spatial information. In scene classification, layout of the image is significant which is related to spatial information. In this paper, we combine spatial pyramid match kernel into CNNs to compensate the spatial information loss and propose spatial ensemble kernel approach to fuse them. By supervised way to learn the kernel parameters in a unit networks combined with CNNs, the whole framework is with increasing discrimination and with compensating layout information.

2. SPATIAL ENSEMBLE KERNEL LEARNING

2.1. Spatial pyramid kernel

In the last few years convolutional neural networks (CNNs) have been the first selection to powerfully represent image contents for various category-level recognition tasks such as object classification [8, 13], scene recognition [14] or object detection [15].

Compared with hand-crafted descriptors, deep learned descriptors can capture the characteristic of image from human visual understanding system. However, traditional CNNs lack spatial information and pay close attention to holistic structure. In order to reveal spatial coherence, spatial pyramid matching (SPM) combined with CNNs is proposed to capture the layout information as shown in Fig. 1.

Three different granularities are adopted in SPM and the whole image is divided into $\{1, 4, 16\}$ grids separately. Each grid is fed into CNNs, which is VGG-16 model pre-learned by ImageNet.

In traditional spatial pyramid match kernel, the final kernel matrix is composed by each grid kernel function as

$$\kappa(x_1, x_2) = \sum_{l=1}^{L} \sum_{i=1}^{I} \kappa(x_1^l(i), x_2^l(i)) \tag{1}$$

where l means different level and i is the index of grid on each level. $\kappa(x_1^l(i), x_2^l(i))$ is the histogram intersection function or inner product. In this paper, we combine them with *Fourier Feature Embedding* in next subsection.

2.2. Fourier Feature Embedding

Kernel approaches are powerful to handle the non-linear relationship embedded in the dataset by $\phi(\cdot)$ to map the low dimension feature space into a high dimensional or even infinite-dimensional feature space. Since $\kappa(x_1, x_2) =< \phi(x_1), \phi(x_2) >$ in most cases, it is no need to explicitly define the mapping function $\phi(\cdot)$. Relying on the implicit lifting provided by the kernel trick, soundable performances for most tasks are ensured.

While with the increasing of dataset scale and considering of the calculating complexity, it is desired explicitly mapping the data to a low-dimensional Euclidean inner product space using a randomized feature map as

$$\kappa(x_1, x_2) =< \phi(x_1), \phi(x_2) >\approx \Phi(x_1)^T \Phi(x_2) \tag{2}$$

Unlike the kernels lifting $\phi(\cdot)$, $\Phi(\cdot)$ is low-dimensional. Thus, we can simply transform the input with $\Phi(\cdot)$, and then apply fast linear learning methods to approximate the answer of the corresponding nonlinear kernel machine.

Kernel approximation is to find explicit $\Phi(\cdot)$.

Theorem 1 *(**Bochner** [16]) A continuous function $g : \mathbb{R}^d \to \mathbb{C}$ is positive definite on $\mathbb{R}^d$ if only if it is the Fourier transformation of a finite non-negative Borel measurement $\mu(\boldsymbol{\omega})$ on $\mathbb{R}^d$, i.e.,*

$$g(\mathbf{x}) = \int_{\mathbb{R}^d} e^{-j\boldsymbol{\omega}^\top \mathbf{x}} d\mu(\boldsymbol{\omega}), \quad \forall x \in \mathbb{R}^d \tag{3}$$

where j denotes the imaginary unit.

Proposition 1 *(**Shift-invariant kernel**) A kernel function $\kappa : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{C}$ is called shift-invariant if $\kappa(x_1, x_2) = k(x_1 - x_2)$ for positive definite function $k : \mathbb{R}^d \to \mathbb{C}$.*

By combining Theorem 1 and Proposition 1, we can get

$$\begin{aligned} \kappa(x_1, x_2) = k(x_1 - x_2) &= \int_{\mathbb{R}^d} e^{j\omega^T (x_1 - x_2)} d\mu(\omega) \\ &= \int_{\mathbb{R}^d} \xi_\omega(x_1) \overline{\xi_\omega(x_2)} d\mu(\omega) \end{aligned} \tag{4}$$

where $\xi_\omega(x) = e^{j\omega^T x}$.

It can be drawn that $\xi_\omega(x_1)\overline{\xi_\omega(x_2)}$ is an *unbiased estimate* of $k(x_1 - x_2)$. Then Eq. (4) is rewritten as

$$\kappa(x_1, x_2) \approx \xi_\omega(x_1)\overline{\xi_\omega(x_2)} = \Phi(x_1)^T \Phi(x_2) \tag{5}$$

where $\Phi(\cdot) = \sqrt{(2/M)}(\xi_{\omega_1}(\cdot), \cdots, \xi_{\omega_M}(\cdot))$.

In order to avoid the complex computing, we can use $z_{\omega,b}(x)$ to replace $\xi_\omega(x)$ as

$$z_{\omega,b}(x) = \sqrt{2}\cos(\omega^T x + b) \tag{6}$$

and

$$\Phi(x) = \sqrt{(2/M)}(z_{\omega_1,b_1}(x), \cdots, z_{\omega_M,b_M}(x)) \tag{7}$$

$[\omega_1, \cdots, \omega_M]$ can be randomly extracted as i.i.d samples from probability distribution $\mu(\omega)$ [17], or by Monte Carlo sampling method [18]. However, [19] views the mapping

function $z_{\omega,b}(x)$ as a neuron with a cosine activation function where biases are as uniform distribution in $[0, 2\pi)$. It built a 2-layer network that can be trained by backpropagation and stochastic gradient descent.

Inspired by this idea, we further combine it into CNNs and spatial pyramid kernel to solve the scene classification problem.

2.3. Ensemble Kernel Learning

In Fig.1, we combines different grid branch in the spatial pyramid by kernels, which is realized by a new nonlinear layer in the neural network with the cosine activations. The parameters of the nonlinear layer can be learned jointly in one single optimization framework by supervised learning, which enables compact and discriminative feature representations.

As shown in Fig.1, by fully connected layer (*fc*), CNNs are extended by two layers as *fcos* layer and *softmax* layer. *softmax* layer just acts as a classifier in our scene classification task, while *fcos* layer conducts combination of different branches by kernel approximation approach. After *fcos* layer, all branches are cascaded together and fed into *softmax* layer. By sharing W_1 matrix, the number of unknown parameters can be reduced.

Supposing dimension of *fc* layer being 4096, and M as the neuron number of one subpart in *fcos* layer, the unknown parameters are only existed in $W_1^{4096\times M}$ and $W_2^{(M\times N)\times num_{class}}$, where $N = \sum_{l=1}^{L} 2^{l-1}$ and num_{class} is the neuron number in *softmax* layer.

In our scene classification task, parameters in CNNs are fixed as pre-trained VGG-16 model with the consideration of training data size. However W_1 and W_2 are learned by a supervised way, which enhances the performance to a great extent. In fact, Fig.1 is an end-to-end model and in case that condition permitted, we can apply an unified loss function to optimize all parameters at one step, which will improve performance.

The loss function during optimization procedure is selected as cross entropy function as

$$L = \sum -\log(\frac{\exp(W_2^T\hat{x} + b_2)}{\sum \exp(W_2^T\hat{x} + b_2)}) \tag{8}$$

where $\hat{x}$ is the cascade of $\hat{x}^l(i)$ over grid index i and pyramid level l. $\hat{x}^l(i)$ is computed as

$$\hat{x}^l(i) = \Phi(x^l(i)) = \frac{2}{\sqrt{M}} cos(W_1^T \times x^l(i)) \tag{9}$$

where the bias b in Eq.(6) is set to zero.

In original spatial pyramid match approach, as shown in Eq.(1), kernel matrix is directly combined over grid and over level. If the $\kappa(x_1^l(i), x_2^l(i))$ is the inner product, then Eq.(1) is changed into

$$\kappa(x_1, x_2) = \sum_{l=1}^{L}\sum_{i=1}^{I} x_1^l(i)^T x_2^l(i) = x_1^T x_2 \tag{10}$$

where x_1 and x_2 are directly cascaded by $x_1^l(i)$ and $x_2^l(i)$ over i and l.

In order to dig the essence of our networks, in Fig.2, we give another equivalent form of the right part networks in Fig.1.

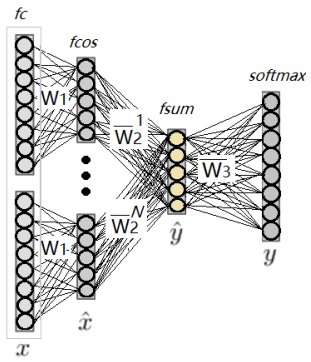

Fig. 2. Equalization structure of right part in Fig.1

Fig.2 adds an additional layer *fsum* with no bias (in yellow color) which keeps the same neuron number as M. Then it is as

$$\begin{aligned} y &= \overline{W}_3^T(\sum_{n=1}^{N}(\overline{W}_2^n)^T\hat{x}^n) + \bar{b}_3 = \overline{W}_3^T\left(\begin{bmatrix} \overline{W}_2^1 \\ \overline{W}_2^2 \\ \vdots \\ \overline{W}_2^N \end{bmatrix}^T \hat{x} + \bar{b}_2\right) + \bar{b}_3 \\ &= \begin{bmatrix} \overline{W}_2^1\overline{W}_3 \\ \overline{W}_2^2\overline{W}_3 \\ \vdots \\ \overline{W}_2^N\overline{W}_3 \end{bmatrix}^T \hat{x} + \overline{W}_3^T\bar{b}_2 + \bar{b}_3 \end{aligned} \tag{11}$$

It can be drawn that Fig.2 keeps the same form as $W_2^T\hat{x} + b_2$ (right part in Fig.1). In addition, from kernel combination aspect, the kernel matrix is as

$$\begin{aligned} \kappa(\hat{y}_1, \hat{y}_2) &= \sum_{l=1}^{L}\sum_{i=1}^{I}\Phi(x_1^l(i))^T \sum_{l=1}^{L}\sum_{i=1}^{I}\Phi(x_2^l(i)) \\ &= \Phi(x_1)^T\Phi(x_2) \end{aligned} \tag{12}$$

where $\Phi(\cdot)$ is as Eq.(7) and is realized by *fcos* layer.

3. EXPERIMENTS AND ANALYSIS

We conduct extensive experiments on two widely used benchmark datasets for scene recognition. We have also compared to state-of-the-art methods to show the great effectiveness of the propose SEK for scene classification.

Table 1. Accuracy (%) with different M on two datasets

M	MIT indoor	SUN 397
100	68.95	50.32
200	75.73	56.58
300	72.66	54.65

3.1. Experimental setup and datasets

After fully connected layer in Fig.1, descriptor of each grid i on each level l is Z-score standardization along dimension as follow.

$$\tilde{x}_d^l(i) = \frac{((x_d^l(i)) - \mu_{x^l(i)})}{\sigma_{x^l(i)}} \tag{13}$$

where d means dimension index, and $\mu_{x^l(i)}$ and $\sigma_{x^l(i)}$ are mean and variance respectively.

We report results on two publicly available datasets, MIT indoor and SUN 397. The training and test separation follows the original setting in each dataset, and for SUN 397, the final results are achieved by 50 training images for each category.

- MIT indoor [20]: MIT indoor focuses on indoor scene categories. It includes five big categories as store, home, public spaces, leisure and working place where each one further contain many subclasses as class, hospital et. al in working space. The whole number of categories is 67. The images in the dataset were collected from different sources: online image search tools (Google and Altavista), online photo sharing sites (Flickr) and the LabelMe dataset. The database contains 15,620 images and all images have a minimum resolution of 200 pixels in the smallest axis.

- SUN 397 [21]: SUN (Scene UNderstanding) 397 dataset contains approximate 100,000 images of 397 categories. For each scene category, images were retrieved using WordNet terminology from various search engines on the web [22]. Only color images of 200×200 pixels or larger were kept. Each image was examined to confirm whether or not it fit a detailed, verbal definition for its category. For similar scene categories (e.g. abbey, church, and cathedral) explicit rules were formed to avoid overlapping definitions.

3.2. Performance analysis on two datasets

To investigate the efficacy with different number of neuron of *fcos* layer, we conduct experiments on two datasets and the results are listed in Table 1 with different M.

It can be seen that the performance is affected by M, the number of neuron in *fcos* layer. In fact, appropriate value of M depends on the scale of dataset and the number of nodes on *softmax* layer. In our experiments, the best performances are as 75.73% and 56.58% on MIT indoor and SUN 397 separately, where M is 200.

Table 2. Comparison on MIT indoor.

Method	Accuracy (%)
DeCaF [23]	59.50
MOP-CNN [11]	68.88
fc8-FV [10]	72.86
MFA-FS [24]	81.43
SEK	**75.73**

Table 3. Comparison on SUN 397 dataset.

Method	Accuracy (%)
Combined 12 feature types [21]	38.00
FV (SIFT) [25]	43.30
DeCaF [23]	43.76
FV (SIFT+LCS) [25]	47.20
MOP-CNN [11]	51.98
fc8-FV [10]	54.40
MFA-FS [24]	63.31
SEK	**56.58**

3.3. Comparison to the state of the art

A comparison on MIT indoor of our proposed approach with other leading representations derived from CNNs is shown in Table 2. Additionally, similar comparison on SUN 397 is listed in Table 3, which also includes hand-crafted representation like SIFT plus Fisher vector [25].

From Table 2 and Table 3, it can be seen that our performances on both datasets are competitive to state-of-the-art algorithms. The main reason of the soundable performance in scene classification lies in the layout information captured by spatial pyramid and ensemble kernel combination, which are both fused with trainditional CNNs.

4. CONCLUSION

In this paper we discussed the benefits of ensemble kernel combination approach by extension CNNs of a new layer as *fcos* with cosine active function. Specially, since layout information is significant for scene classification, spatial pyramid is applied in our whole framework. We named it as spatial ensemble kernel learning approach, which can unite different kind of information by kernel. Combined with CNNs, it can further be treated as an end-to-end networks. Experiments on MIT indoor and SUN 397 datasets proved the effectiveness of proposed approach.

5. REFERENCES

[1] Colin Mcmanus, Winston Churchill, Will Maddern, Alexander D. Stewart, and Paul Newman, "Shady dealings: Robust, long-term visual localisation using illumination invariance," in *IEEE International Conference on Robotics and Automation*, 2014, pp. 901–906.

[2] Sven Middelberg, Torsten Sattler, Ole Untzelmann, and Leif Kobbelt, "Scalable 6-dof localization on mobile devices," in *European Conference on Computer Vision*, 2014, pp. 268–283.

[3] Mathieu Aubry, Bryan C. Russell, and Josef Sivic, *Painting-to-3D model alignment via discriminative visual elements*, ACM, 2014.

[4] James Philbin, Ondrej Chum, Michael Isard, Josef Sivic, and Andrew Zisserman, "Object retrieval with large vocabularies and fast spatial matching," in *Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on*, 2007, pp. 1–8.

[5] Relja Arandjelovic and Andrew Zisserman, "All about vlad," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 1578–1585.

[6] F. Perronnin, Y. Liu, J. Snchez, and H. Poirier, "Large-scale image retrieval with compressed fisher vectors," in *Computer Vision and Pattern Recognition*, 2010, pp. 3384–3391.

[7] Mark Everingham, S. M. Ali Eslami, Luc Van Gool, Christopher K. I. Williams, John Winn, and Andrew Zisserman, "The pascal visual object classes challenge: A retrospective," *International Journal of Computer Vision*, vol. 111, no. 1, pp. 98–136, 2015.

[8] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton, "Imagenet classification with deep convolutional neural networks," in *International Conference on Neural Information Processing Systems*, 2012, pp. 1097–1105.

[9] Relja Arandjelovic, Petr Gronat, Akihiko Torii, Tomas Pajdla, and Josef Sivic, "Netvlad: Cnn architecture for weakly supervised place recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 5297–5307.

[10] Mandar Dixit, Si Chen, Dashan Gao, Nikhil Rasiwasia, and Nuno Vasconcelos, "Scene classification with semantic fisher vectors," in *Computer Vision and Pattern Recognition*, 2015, pp. 2974–2983.

[11] Yunchao Gong, Liwei Wang, Ruiqi Guo, and Svetlana Lazebnik, "Multi-scale orderless pooling of deep convolutional activation features," in *European Conference on Computer Vision*, 2014, pp. 392–407.

[12] Bolei Zhou, Agata Lapedriza1, Jianxiong Xiao, Antonio Torralba, and Aude Oliva, "Learning deep features for scene recognition using places database," in *Advances in Neural Information Processing Systems*, 2014, pp. 1–9.

[13] M Oquab, L Bottou, I Laptev, and J Sivic, "Learning and transferring mid-level image representations using convolutional neural networks," in *Computer Vision and Pattern Recognition*, 2014, pp. 1717–1724.

[14] Bolei Zhou, Agata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva, "Learning deep features for scene recognition using places database," in *International Conference on Neural Information Processing Systems*, 2014, pp. 487–495.

[15] Ross B. Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Computer Vision and Pattern Recognition*, 2014, pp. 1–8.

[16] Walter Rudin, *Fourier analysis on groups*, John Wiley & Sons, 2011.

[17] Ali Rahimi and Benjamin Recht, "Random features for large-scale kernel machines," in *Neural Infomration Processing Systems*, 2007, pp. 1177–1184.

[18] Felix X Yu, Sanjiv Kumar, Henry Rowley, and Shih Fu Chang, "Compact nonlinear maps and circulant extensions," *Computer Science*, 2015.

[19] John Moeller, Vivek Srikumar, Sarathkrishna Swaminathan, Suresh Venkatasubramanian, and Dustin Webb, "Continuous kernel learning," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 2016, pp. 657–673.

[20] A. Quattoni and A. Torralba, "Recognizing indoor scenes," pp. 413–420, 2001.

[21] Jianxiong Xiao, James Hays, Krista A. Ehinger, Aude Oliva, and Antonio Torralba, "Sun database: Large-scale scene recognition from abbey to zoo," in *Computer Vision and Pattern Recognition*, 2010, pp. 3485–3492.

[22] Torralba A, Fergus R, and Freeman WT, *80 million tiny images: a large data set for nonparametric object and scene recognition*, vol. 33, IEEE Trans. on Pattern Analysis and Machine Intelligence, 2008.

[23] 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," vol. 50, no. 1, pp. I–647, 2013.

[24] Mandar Dixit and Nuno Vasconcelos, "Object based scene representations using fisher scores of local subspace projections," in *Advances in Neural Information Processing Systems*, 2016.

[25] Jorge Snchez, Florent Perronnin, Thomas Mensink, and Jakob Verbeek, "Image classification with the fisher vector: Theory and practice," *International Journal of Computer Vision*, vol. 105, no. 3, pp. 222–245, 2013.