New Deep Kernel Learning based Models for Image Classification

Rabha O.Abd-elsalam*, Yasser F.Hassan*, Mohamed W.Saleh*
*Department of Mathematics and computer science
Faculty of science, Alexandria University
Alexandria, Egypt

***Abstract*—Deep learning system is used for solving many problems in different domains but it gives an over-fitting risk when richer representations are increased. In this paper, three different models with different deep multiple kernel learning architectures are proposed and evaluated for the breast cancer classification problem. Discrete Wavelet transform and edge histogram descriptor are used to extract the image features. For image classification purpose, support vector machine with the proposed deep multiple kernel models are used. Also, the span bound is employed for optimizing these models over the dual objective function. Furthermore, the comparison between the performance of the traditional support vector machine which uses only single kernel and the introduced models is worked out that show the efficiency of the experimental results of the proposed models.**

Keywords—Deep learning; multiple kernel; support vector machine; image classification

I. INTRODUCTION

Recently, deep learning techniques are used for solving many problems in different domains as a result of performing well when training the regression model in high-dimensional data. Deep learning techniques succeed in both, machine learning and traditional computer vision. But, the identification of the application condition is needed for the deep learning. Many researchers make different studies to discover the pros and cons of deep learning over other machine learning methods. The most direct form is making a comparison between deep learning architectures and support vector machine (SVM) in processing audio, images and videos. However, there are not enough studies to choose the parameters once using deep learning technique for regression and classification tasks [1]-[5].

In machine learning filed, kernel learning technique is an active research subject [6], [7]. Kernel principal component analysis (KPCA) and SVM are the most common methods relay on kernel techniques. These kernel approaches have been applied to different applications due to their good performance. Unfortunately, the performance of those approaches depends on the selected kernel [8-10]. Thus, different studies have been introduced to learn the best kernel for these approaches [11], [12].

Multiple kernel learning (MKL) has been suggested to state the limits of single fixed kernel techniques. Bach et al. introduced the first MKL formulation [11]. Recently, MKL has been developed for automated kernel parameter tuning. Its goal is to learn a linear or convex combination of multiple regular kernels to define the best target kernel for the given application [13], [14].

Many algorithms for extended MKL methods have been introduced to enhance the performance of the regular MKL method. In some real applications, MKL methods do not always yield better experimental performance once they compared with the regular techniques. Therefore, the deep learning architectures [15]-[17] are very promising choices than the shallow one. Furthermore, they can be used for feature extraction and in kernel applications as classifier (multilayer of multiple kernels learning (MLMKL)) [10].

The authors in [18] introduced a novel kernel which mimics the deep learning structure. They obtained static network where fixed kernels are used without learning the optimal kernels combination. A general framework for MLMKL is proposed in [10]. The authors had some problems in network optimization beyond two layers. The second layer only contains a single radial basis function (RBF) kernel.

The authors in [19] optimized the MLMKL with several layers and they used the leave-one-out error estimation algorithm. Unfortunately, their method is not evaluated over the MKL. Furthermore, no enhancements were achieved when using more than two layers.

In this paper, three models for deep kernel learning (DKL) are proposed and evaluated in the breast cancer classification problem. Additionally, span bound is exploited for the sake of optimizing the proposed models over the dual objective function. A comparison between the performance of the regular SVM using single kernel and the proposed DKL models is introduced.

The paper is organized as follows. The multi-layer multiple kernel deep learning is briefly described in Section 2. While Sections 3 and 4, introduce the methodology and the proposed deep kernel models. The experimental results are explained in Section 5. Section 6 concludes the work and presents the future work.

II. MULTI-LAYER MULTIPLE KERNEL DEEP LEARNING

A. Multiple Kernel Learning

Suppose that $\{(a_1, b_1), \dots, (a_m, b_m)\}$ are m training samples where $a_j \in \mathbb{R}^d$ is the feature vector of the sample and b_j is the sample label. The problem of MKL is generally described as the constrained optimization problem [10], [11]:

$$min \quad \lambda\|f\|_{H_k} + \sum_{j=1}^{m} \ell\left(b_j f(a_j)\right), \qquad (1)$$
$$s.t \qquad k \in \mathcal{K} \ , \ f \in H_k$$

Where $\ell(\cdot)$ refers to some loss function like $\ell(u) = max(0, 1-u)$ that used in SVM, λ is the regularization parameter, $\mathcal{K}$ represents the candidate kernels optimization domain, and H_k is the reproducing kernel Hilbert space related to the k kernel.

In (1), the decision function $f(a)$ is in the form of linear expansion of kernel evaluation on the training samples a_j,

$$f(a) = \sum_{j=1}^{m} \alpha_j \ k(a_j, a) \qquad (2)$$

Where α_j are the coefficients referred to in [10].

In [10], the kernel is a set of convex combination of predefined base kernels:

$$\mathcal{K}_{conv} := \left\{ k(.,.) = \sum_{i=1}^{n} \rho_i \ k_i(.,.): \sum_{i=1}^{n} \rho_i = 1 \ , \quad \rho_i \geq 0, \quad i = 1, \dots, n \right\}, \qquad (3)$$

where each candidate k is the summation of the n base kernels $\{k_1, \dots, k_n\}$, and ρ_i is the coefficient of the i^{th} base kernel. So, the decision function can be expanded with the multiple kernels as:

$$f(a) = \sum_{j=1}^{m} \alpha_j \sum_{i=1}^{n} \rho_i \ k_i(a_j \ , a) = \sum_{j=1}^{m} \sum_{i=1}^{n} \alpha_j \ \rho_i k_i(a_j, a), \qquad (4)$$

and the last kernel will be a linear summation of $\boldsymbol{n}$ base kernels.

B. Deep Kernel Learning

Recently, many studies show that there is a limitation in conventional learning methods concerning their learning structural design. The deep structural design is often better than the shallow ones. The idea of deep learning of kernel methods that introduced in [19], [20] can be applied either in shallow structures such as SVM or in deep architectures.

The l - layer kernel is the inner product after several feature mapping of inputs:

$$k^{(l)}(a_i, a_j) = \langle \underbrace{\varphi(\varphi(\dots(\varphi(a_i))))}_{l-\ times}, \underbrace{\varphi(\varphi(\dots(\varphi(a_j))))}_{l-\ times} \rangle \qquad (5)$$

Here φ is the essential feature mapping function of k and $\langle \cdot, \cdot \rangle$ represents the inner product.

Polynomial kernel is considered as an example of two-layer kernel, such as:

$$k^{(1)}(a \ , b) = \ (\delta(a.b) + \ \gamma)^d$$

$$k^{(2)}(a \ , b) = \ \left(\delta\left(k^{(1)}(a, b)\right) + \ \gamma\right)^d, \qquad (6)$$

Where δ , γ and d refer to the free parameters of the polynomial kernel. The Gaussian RBF kernel can be approximated as:

$$k^{(1)}(a, b) \ \approx \ k^{(2)}(a, b) = \ \varphi^{(2)}\left(\varphi^{(1)}(a)\right).\varphi^{(2)}\left(\varphi^{(1)}(b)\right) = \ e^{-2\lambda\left(1-k(a.b)\right)}. \qquad (7)$$

The DKL has been suggested to use the deep learning idea for improving the MKL task.

A domain of l - level multi-layer kernels is defined as follows:

$$k^{(l)} = \left\{ k^{(l)}(.,.) = \boldsymbol{q}^{(l)}(\left[k_1^{(l-1)}(.,.), \dots, k_n^{(l-1)}(.,.)\right]) \right\} \qquad (8)$$

Where $q^{(l)}$ is a function to pool multiple $(l-1)$ level kernels that should guarantee the valid resulting kernel.

The optimization problem of l- level MLMKL is described as:

$$min_{k \in k^{(l)}} \, min_{f \in H_k} \, \lambda\|f\|_{H_k} + \sum_{j=1}^{m} \ell(b_j f(a_j)) \qquad (9)$$

III. METHODOLOGY

The design of the image recognition system generally involves collection data, feature extraction, model selection or training, and evaluation. This part describes the design of the recognition system for the breast cancer classification problem in the digital image.

A. Data Collection

The breast cancer databases are sets of mammograms images. This work used BCDR-F01 (Film Mammography dataset number 1) which is the first dataset of BCDR. The BCDR-F01 is a binary class dataset which composed by biopsy (Benign vs. Malign) [21].

B. Features Extraction

Feature descriptors play an important role in recognition system. Really, they permit a mapping from visual information to a numerical vector which returns the semantic contents of the images. Regarding features extraction, this work used MPEG-7 edge histogram descriptor (EHD) [22] as input to train, evaluate and compare the proposed models and the traditional SVM classifier. EHD is used to refer the frequency and directionality of edges within each image region. Initially, simple edge detector operator is used for identifying edges and grouping them into five categories: horizontal, vertical, diagonal, anti-diagonal and non-edge. Then, global, local and semi-local edge histograms are calculated. The EHD features are represented by a vector of diminution 150.

Additionally, this work used discrete wavelet transform (DWT) to decompose an input digital image into four sub-bands of different frequencies [23]. The four sub-bands are generally denoted as approximation image (LL), horizontal (HL), vertical (LH) and diagonal (HH) detail components. The LL sub-band is used in this experiment to hold the most useful information of the input image.

C. Classification

This work employed SVM for breast cancer classification, which is a two-class problem. SVM is a machine learning method that involves training and testing steps. With the two-class problem, training samples (a_j, b_j) are given, where $a_j \in \mathbb{R}^d$ is the feature vector of the given sample and b_j is the label of its class, (+1 and -1 point to the two classes which are benign and malign classes respectively). SVM builds an optimal hyper-plane that maximizes the margin to classify the samples [24].

Traditionally, the margin is maximized through the gradient of the dual objective function with respect to the kernel hyper-parameters. But, the structures of deep learning give an over-fitting risk when richer representations are increased. So, looking for a tight bound of the leave-one-out error is needed. This paper used the span bound due to its promising results in single layer multiple kernel learning. The span bound is defined as:

$$T_{span} \coloneqq L\big((a,b),\dots,(a_n,b_n)\big) \leq \sum_{p=1}^{n} \varphi\left(\alpha_p^* S_p^2 - 1\right) \tag{10}$$

Where, L points to the leave-one-out error and S_p refers to the distance between the support vector and the set Λ_p [25], [26] where:

$$\Lambda_p = \left\{ \sum_{i \neq p, \alpha_i^* > 0} \lambda_i \, \boldsymbol{\varphi}_{\boldsymbol{k_\theta}}(\boldsymbol{a_i}) \;\middle|\; \sum_{i \neq p} \lambda_i = 1 \right\}. \tag{11}$$

IV. Proposed Model

It is usually agreed that SVM is highly depend on the selected kernel function. In the regular SVM, the kernel function maps the input data, and then, the SVM is trained using this input data for the classification task. MKL is one probable structure, which designs the multiple kernels as linear combinations of base functions. Instead of using a single kernel function, a set of kernel functions can be organized in a particular structure to transform the original data over a number of layers of kernels. Then, the final kernel is used to learn the SVM decision function. The gradient descent presented in [19] is adopted in this work for optimizing the weights of the proposed deep kernels.

In this paper, three different models with different framework are considered for deep kernel learning architecture as shown in Fig. 1, 2 and 3 where the lines represent the weights for each kernel. Every model tries to optimize the weights of its architecture. The number of kernels in each layer is two in the three models. The first and third models have three layers, while the second model has only two layers.

The first model in Fig. 1 explores the combination of multiple kernels (two kernels). The elementary kernels in the first layer are computed from the given data and fed as input to the deep structure. The final kernel is learned as a three-multi-layered linear combination of functions where each one takes in a combination of two basic or two intermediate functions on multiple features.

In the second model shown in Fig. 2, the first and the second kernels (k_1^1, k_2^1) in the first layer and the first kernel (k_1^2) in the second layer transform the given input data. On the other hand, the second kernel (k_2^2) in the second layer takes the linear combination of the output of the first and second kernels (k_1^1, k_2^1) in the first layer. The final kernel is learned as a linear combination of the output of the two kernels in the second layer (k_1^2, k_2^2).

In the third model shown in Fig. 3, the two kernels of the first layer map the original given data. While the first kernel (k_1^2) in the second layer map the output of the second kernel (k_2^1) in the first layer. The second kernel (k_2^2) in the second layer maps the output of the first kernel (k_1^1) in the first layer. The first kernel (k_1^3) in the third layer maps the combination of the output of the two kernels (k_1^2, k_2^2) in the second layer. While the second kernel (k_2^3) in the third layer maps the combination of the original data and the output of the first kernel (k_1^2) in the second layer. The final kernel is a combination of the output of the kernels (k_1^3, k_2^3) in the third layer.

V. Experimental Results

The proposed DKL system has been implemented in MATLAB® 2015b with Windows 7 enterprise edition environment. The BCDR-F01 dataset is used to test and evaluate the performance of the DKL system in breast cancer classification problem. BCDR-F01 is a binary class dataset which composed by biopsy (Benign vs. Malign) [21].

The proposed models are tested on 86 images (29 benign images and 57 malignant images). The label of the benign class is +1 while the label of the malign class is -1. In the classification operation, 50% of images are used for training the classifier and 50% for testing the trained classifier; the images are randomly selected for training and testing stages. For the sake of comparing the performance of the proposed DKL models and the regular SVM, the performance is given in terms of accuracy which is the proportion of the correct classified samples to the total number of samples.

$$Accuracy = \frac{TP + TN}{N} \tag{12}$$

Where, TP is the number of true positive, TN is the number of true negative, and N is the total number of instances in the test set.

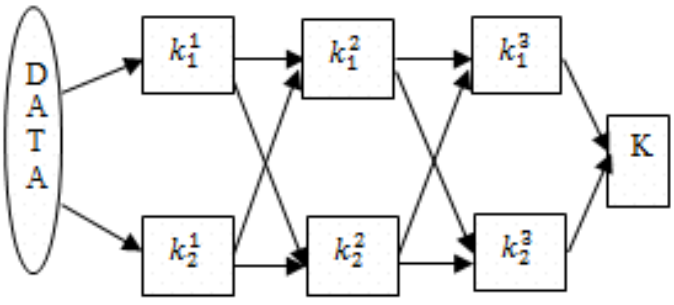

Fig. 1. DKL architecture for three layers with two kernels in each layer.

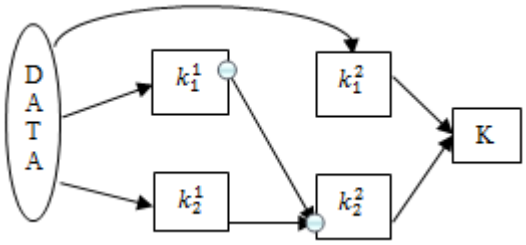

Fig. 2. DKL architecture for two layers with two kernels in each layer.

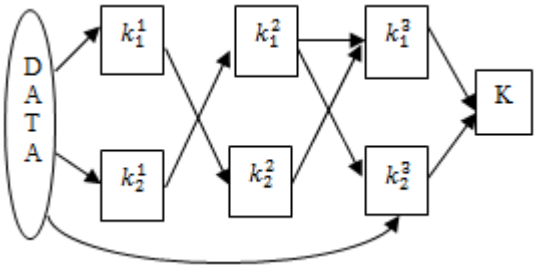

Fig. 3. DKL architecture for three layers with two kernels in each layer.

The same type of kernels in the architecture is used for those models (either all are RBF kernels or polynomial kernels). Multiple Kernels are considered for generating diverse representation of the data with basic functions. Deep learning structures present an over-fitting risk when richer representations are increased [19]. The over-fitting can be avoided by selecting a small number of base kernels. So, two kernels are used in each layer. Furthermore, the span bound is employed for finding a tight bound of the leave-one-out error. Span bound, presented promising results in [19], [26] with single layer multiple kernel learning over the dual objective function. In this paper, the gradient descent is used on the span bound for 100 iterations to DKL structure. The SVM penalty constant is fixed to 10 and the value of the learning rate is 10^{-4}.

Table 1 illustrates the accuracy of the proposed three DKL models and the regular SVM, which uses only single kernel, using the feature extraction methods (EHD and DWT). DWT achieves better accuracy than EHD with the second and third DKL models when RBF kernel is used. But, EHD descriptor gives better accuracy than DWT with all models when polynomial (POLY) kernel is used. The third model achieves better results than other models with the two feature extraction methods.

POLY kernel gives better accuracy than RBF kernel in all models. When the DKL system has been tested with the POLY kernel, the first kernel in each layer is the POLY kernel with the degree of 2 while the second kernel is the POLY kernel with the degree of 5. The POLY kernel with the degree of 2 is already flexible to discriminate between the two classes with a good margin. Also, the POLY kernel with the degree of 5 yields a similar decision boundary. Model 3 achieves the best results among all models due to its deep architecture which can help to boost accuracy as shown in shaded cell in Table 1 (88%).

TABLE. I. A COMPARISON AMONG OUR PROPOSED DKL MODELS AND REGULAR SVM

Kernel Type	**Feature Extraction**	**Accuracy**			
		Classification Method			
		Model1	**Model2**	**Model3**	**SVM**
RBF Kernel	EHD	66.67	64.29	64.29	69.05
	DWT	66.67	76.19	76.19	66.67
Polynomial Kernel	EHD	76.19	85.71	**88.10**	73.81
	DWT	71.43	80.95	83.33	71.43

VI. CONCLUSION AND FUTURE WORK

In this paper, three DKL models for breast cancer classification problem are introduced. Span bound is used for optimizing the proposed models over the dual objective function. A comparison between the performance of the regular SVM which uses only single kernel and the proposed models is introduced. The experimental results show that the proposed models overcome the traditional SVM. Furthermore, model 3 gives the best results among the other models due to its deep architecture that can help boost accuracy.

New features sets with another deep kernel structures will be explored on bigger dataset for the sake of determining which features set is the most discriminative with respect to breast cancer classification problem. Since the DKL has a bigger adaptability to data (because it's based on the creation of an optimal kernel to fit that data). These orientations will be the ultimate subject of the future work.

REFERENCES

[1] P. Liu, K.-K.R. Choo, L. Wang, F. Huang, Soft Computing (2016) 1-13.

[2] G. Omer, O. Mutanga, E.M. Abdel-Rahman, E. Adam, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 8 (2015) , pp. 4825-4840.

[3] Y. LeCun, Y. Bengio and G. Hinton. "Deep Learning", Nature 521 (2015) , pp. 436-444.

[4] W. Zhao, S. Du. "Spectral–spatial feature extraction for hyper spectral image classification: A dimension reduction and deep learning approach." IEEE Transactions on Geoscience and Remote Sensing 54 (2016) , pp. 4544-4554.

[5] J.J. Tompson, A. Jain, Y. LeCun, C. Bregler, Joint training of a convolutional network and a graphical model for human pose estimation, Advances in neural information processing systems, 2014, pp. 1799-1807.

[6] I. Rebai and Y. Belayed and W. Mahdi. "Deep multilayer multiple kernel learning". Neural Computing and Applications 27 (2016) , pp. 2305-2314.

[7] J. Shawe-Taylor, N. Cristianini, Kernel methods for pattern analysis, Cambridge university press, 2004.

[8] M. Wiering, M. Van der Ree, M. Embrechts, M. Stollenga, A. Meijster, A. Nolte, L. Schomaker, The neural support vector machine, BNAIC 2013: Proceedings of the 25th Benelux Conference on Artificial

Intelligence, Delft, The Netherlands, November 7-8, 2013, Delft University of Technology (TU Delft); under the auspices of the Benelux Association for Artificial Intelligence (BNVKI) and the Dutch Research School for Information and Knowledge Systems (SIKS), 2013.

[9] X. Xu, I. W. Tsang and D. Xu. "Soft margin multiple kernel learning". IEEE Trans Neural Netw Learn Syst 24 (2013) 749-761.

[10] J. Zhuang, I.W. Tsang, S.C. Hoi, Two-Layer Multiple Kernel Learning, AISTATS, 2011, pp. 909-917.

[11] F.R. Bach, G.R. Lanckriet, M.I. Jordan, Multiple kernel learning, conic duality, and the SMO algorithm, Proceedings of the twenty-first international conference on Machine learning, ACM, 2004, p. 6.

[12] M. Hu, Y. Chen and J. T.-Y. Kwok, "Building sparse multiple kernel SVM classifiers," IEEE Transactions on Neural Networks 20 (2009) 827-839.

[13] M. Jiu, H. Sahbi, Deep kernel map networks for image annotation, Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on, IEEE, 2016, pp. 1571-1575.

[14] Z. Xu, R. Jin, I. King, M. Lyu, An extended level method for efficient multiple kernel learning, Advances in neural information processing systems, 2009, pp. 1825-1832.

[15] G. Chen, C. Parada, G. Heigold, Small-footprint keyword spotting using deep neural networks, Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on, IEEE, 2014, pp. 4087-4091.

[16] D. C. Cireşan, U. Meier, L. M. Gambardella and J. Schmidhuber, "Deep, big, simple neural nets for handwritten digit recognition", Neural Computation 22 (2010) 3207-3220.

[17] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, Advances in neural information processing systems, 2012, pp. 1097-1105.

[18] Y. Cho, L.K. Saul, Kernel methods for deep learning, Advances in neural information processing systems, 2009, pp. 342-350.

[19] E.V. Strobl, S. Visweswaran, Deep multiple kernel learning, Machine Learning and Applications (ICMLA), 2013 12th International Conference on, IEEE, 2013, pp. 414-417.

[20] Z. Sun, N. Ampornpunt, M. Varma, S. Vishwanathan, Multiple kernel learning and the SMO algorithm, Advances in neural information processing systems, 2010, pp. 2361-2369.

[21] BCDR . "Breast Cancer Digital Repository," http://bcdr.inegi.up.pt, April 4, 2016.

[22] N. Prajapati and A. K. Nandanwar. "Edge Histogram Descriptor Geometric Moment and Sobel Edge Detector Combined Features Based Object Recognition and Retrieval System," (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 7 (1) (2016) 407-412.

[23] A. Goel. "Discrete Wavelet Transform (DWT) with Two Channel Filter Bank and Decoding in Image Texture Analysis." International Journal of Science and Research (IJSR) Volume 3 (April-2014).

[24] P. Liu, K.-K. R. Choo, L. Wang et al., "SVM or deep learning? A comparative study on remote sensing image classification," Soft Computing, 2016, p. 1-13.

[25] B. Ghattas, A. B. Ishak, "An Efficient Method for Variables Selection Using SVM-Based Criteria." , (2005).

[26] Y. Liu, S. Liao, Y. Hou, Learning kernels with upper bounds of leave-one-out error, Proceedings of the 20th ACM international conference on Information and knowledge management, ACM, 2011, pp. 2205-2208.