Capsule Endoscopy Image Classification with Deep Convolutional Neural Networks

Yaxing Cao, Wenming Yang*, Kaiquan Chen, Yong Ren, Qingmin Liao
Shenzhen Key Lab. of Info. Sci&Tech/Shenzhen Engineering Lab. of IS&DCP,
Department of Electronic Engineering/Graduate School at Shenzhen, Tsinghua University, China
e-mail: *Corresponding author, yangelwm@163.com

***Abstract*—Wireless capsule endoscopy (WCE) is quite an advanced, patient-friendly and novel medical equipment for non-invasive gastrointestinal disease detection, which is able to view the entire gastrointestinal tract without pain. Since there are more than 80,000 WCE images for each examination, it usually takes hours for professional clinicians to diagnose all the video data. Therefore, the automatic computer-aided lesion classification technique is highly perspective needed. In this paper, we propose an effective scheme to classify different lesion images acquired by WCE. Firstly, we obtain feature maps of the same resolution by performing a max pooling operation on different convolutional layers, and then quantify the pooled feature maps by trainable weight parameters, and finally one by one convolution kernels are employed to merge the combined quantized feature maps. We enhance the performance of feature extraction by merging multi-level convolutional features, including both low level and high level features. The preliminary experimental result shows that the classification accuracy rate is up to 95.15%, both running speed and recognition rate are higher than traditional machine learning algorithms.**

Keywords-wireless capsule endoscopy; feature extraction; lesion classification; deep convolutional neural network

I. INTRODUCTION

Traditional endoscopic image recognition algorithms usually extract designed features from lesion areas. Many representative hand-crafted image operators, like Histogram of Oriented Gradient (HOG) and Local Binary Pattern (LBP), are introduced to describe edge features or local texture features of images. Based on these texture feature descriptor operators, machine learning methods such as support vector machine, back-propagation neural networks, adaboost and some other methods are employed to classify the image data.

As for wireless capsule endoscopy (WCE) image lesion recognition, there has already existed numerous research methods. [1] extracts each WCE frame's features based on HSV color histogram and proposes a block edge directivity descriptor. Through experiments, a fixed threshold is selected to cut the WCE video into different clips. Lastly, key frames are extracted using a relation matrix rank method. [2] fuses HSV color features, LBP texture features and HOG shape features based on the information entropy. By comparing the distance of information entropies between two consecutive WCE frames with an automatic threshold, the original WCE video is segmented into several clips. For each clip, AP clustering method is adopted to select key frames. [3] takes both HSV color histogram-based color features and GLCM-based texture features into consideration. A W-parametric mean value threshold is set adaptively to judge the similarity between adjacent WCE frames sequentially for clip segmentation. Eventually, in each clip, the key frames are extracted by an adaptive K-means clustering algorithm. In general, all these methods discussed above are fairly effective to meet their specific demands. However, considering the high requirements for accuracy and efficiency in medical diagnosis, it is necessary to develop more effective WCE classification methods.

In 2012, deep convolutional neural network (DCNN) made a great achievement in the ILSVRC-2012 race. Since then, DCNN has caught many scholars' attention and been applied to other relevant fields quickly, such as object detection, face recognition, handwriting recognition and so on. Currently, DCNN poses a challenge to many advanced algorithms and also shows better performance in pattern recognition and image classification [6].

In this paper, we propose a high-performance algorithm based on DCNN to classify WCE images into five kinds of lesions, including normal, bleeding, ileal erosion, colitis and gastritis. Furthermore, we perform preliminary experiments on the WCE image database. The result shows that our proposed algorithm outperforms SVM-LBP [4] and BP-Contourlet [5]. At the same time, the recognition accuracy rate is up to 95.15%.

The rest of the paper is organized as follows. Section II introduces convolutional neural networks (CNN) in brief and describes our proposed algorithm in detail. In Section III, we conduct extensive experiments to compare our proposed scheme with conventional machine learning methods and two CNN-based image classification schemes. Finally, we summarize our work in Section IV.

II. DETAILS OF PROPOSED METHOD

When designing CNN, both the generality of CNN and the optimization of special applications need to be taken into account. In the actual design work, we should pay more attention to both expansibility and real-time.

In most deep learning algorithms, the depth of network is an important parameter [6] [11], and the DCNN is no exception. VGG nets [10] even specifically verified the effect of depth on the result. It uses 3×3 convolution kernels, and tests the results obtained by different network structures of 11, 13, 16 and even 19 layers respectively. Although the

978-1-5386-8339-2/18/$31.00 ©2018 IEEE

training parameters were slightly increased (from 133M to 144M), it was negligible compared with the correct rate.

The advantages of CNN are as follows: Increasing depth does not lead to a remarkable growth of the training time, but a relatively slow growth. And the final results will be significantly improved with the augment of the network's depth and width. CNN uses the shared parameters between the layers, which not only reduces the required memory size and the number of model parameters to be trained, but also improves the performance of the algorithm. At the same time, there is almost no necessary for preprocessing or eigenvalue extraction, which is the distinct advantage that other machine learning algorithms don't have.

The classical CNN [9] includes convolutional neural layer, the rectified linear units layer (hereinafter referred to as ReLU layer), pooling layer and normalized layer, and its structure is shown in Fig.1.

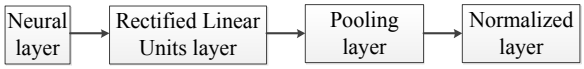

Figure 1. Classical CNN structure.

DCNN model designed in this paper has nine layers in total. The DCNN-based WCE lesion image classification system is illustrated in Fig. 2, which consists of three parts, the convolution layer, pooling layer and three fully-connected layers. The details will be described in the following.

Our proposed scheme has six convolutional layers and each layer consists of one or two contiguous convolutions. The quantized weight parameter W in different layers has corresponding size of $1\times1\times$ channels, which is multiplied by the max pooling operation generating the new feature map x to generate output quantitative feature map y. The input-output model is based on the following equation:

$$y_j = \sum_i W_{ij} \times x_i \qquad (1)$$

where W are trainable model parameters for each newly constructed convolutional layer. Then after getting the quantization feature map y of each layer, all these y are assembled as a new feature map, which is connected to the 'Layer6' to reduce the number of combined feature map channels and also merge low-level and high-level feature maps. The final three fully-connected layers actually function as a classifier, using the features extracted by the previous layers as input to perform the WCE image classification task.

In this design, the system shown in Fig. 2 is trained in a supervised way, where stochastic gradient descent method is employed to minimize the difference between the actual output and desired output at the fully–connected layer. In our research, the gradients are computed using the back propagation (BP) method. All the coefficients of all the filters will be updated together with other parameters in every layer during the training procedure. And the specific model structure is shown in table I. In the experiment, all of the convolution kernel size used is 5×5.

The studied network is trained with stochastic gradient descent method to minimize the weighted combination of the loss functions where the proposed network structure is parameterized by weights W^*. The loss function is shown as below:

$$W^* = \arg\min_W E_{x,\{yi\}}[\sum \lambda_i l_i(f_W(x), yi)] \qquad (2)$$

It transforms the WCE images into corresponding output labels $\tilde{y} = f_W(x)$. The loss function computes a scalar value $l_i(\tilde{y}, yi)$ which measures the difference between the output WCE image label $\tilde{y}$ and the target WCE image label yi. We use a network φ which has been pretrained for image classification as a fixed loss network to define our loss functions.

TABLE I. THE PARAMETERS OF THE DESIGNED DCNN

Layer name		Kernel num	Kernel size	Channel	Padding	Stride
Layer1	Conv1	64	5	3	'SAME'	—
	Pool1	—	2	—	—	2
Layer2	Conv2	128	5	64	'SAME'	—
	Pool2	—	2	—	—	2
Layer3	Conv3_1	256	5	128	'SAME'	—
	Conv3_2	256	5	256	'SAME'	—
	Pool3	—	2	—	—	2
Layer4	Conv4_1	512	5	256	'SAME'	—
	Conv4_2	512	5	512	'SAME'	—
	Pool4	—	2	—	—	2
Layer5	Conv5_1	512	5	512	'SAME'	—
	Conv5_2	512	5	512	'SAME'	—
	Pool5	—	2	—	—	2
Layer6	Conv6	480	1	1472	'SAME'	—
Layer7	Fc1	Size: 3×3×480×1024				
Layer8	Fc2	Size: 1024×512				
Layer9	Fc3	Size: 512×5				
Layer0	Out	—				

For Fig. 2, the inputs are the 5 kinds of WCE image lesions that are scaled into 96×96 from the original size of 256×240. Experiments show that the result of using 5×5 convolution kernel is significantly better than that of 3×3 convolution kernel. The fully-connected layers have 1024, 512 and 5 neurons for each other. The initial weights of each convolutional layer will obey the Gaussian distribution whose mean is zero and standard deviation is 0.01. Then the weights of the fully-connected layers are the same with the convolutional layer except that the standard deviation is set to 0.1. We use the mini-batch stochastic gradient descent method to train our system. The learning rate is initialized to 0.001 and will be modified according to the training times and the batch size. During training, the input to our network is color images with a fixed-size of 96 × 96. The only preprocessing that we do is subtracting the mean of RGB value. Our overall implementation is built on Python3.5.0 and TensorFlow1.4.0.

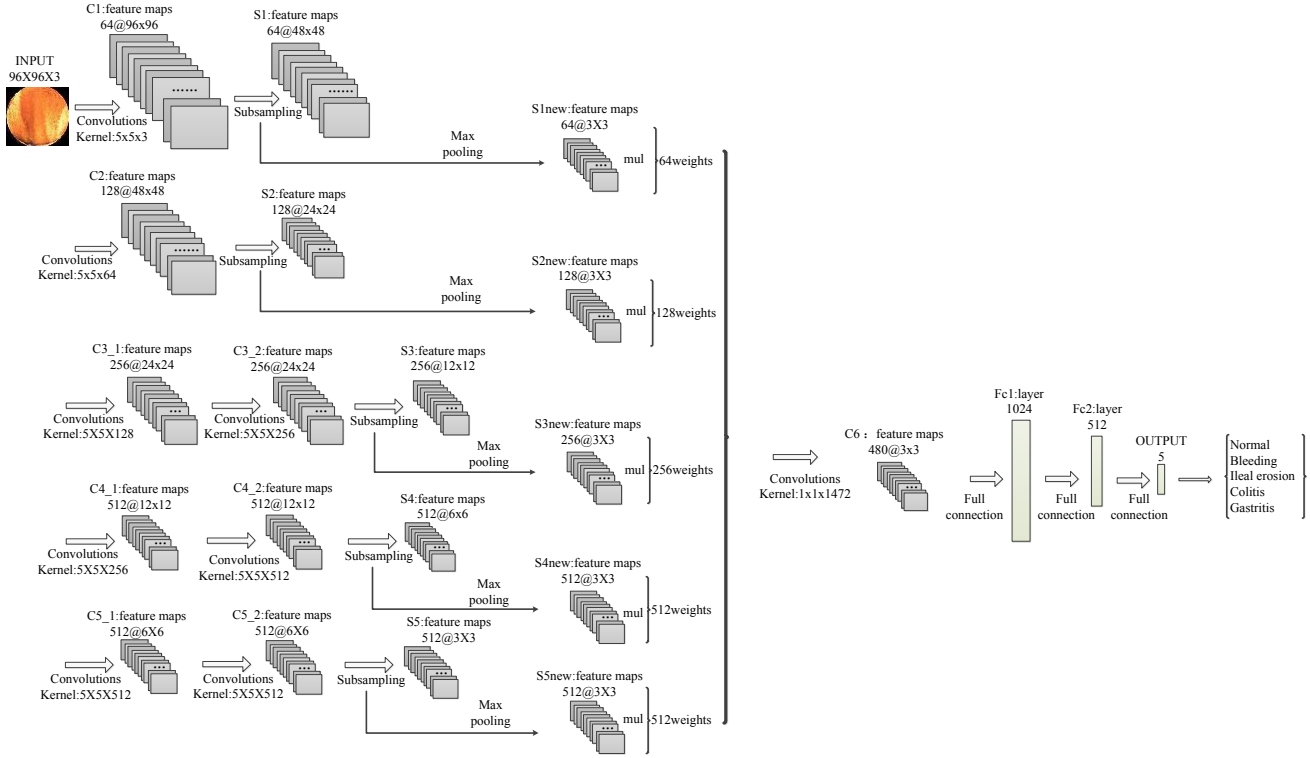

Figure 2. The proposed network structure diagram.

III. EXPERIMENTAL RESULTS

The WCE images used for the experiment are all from real clinical data. We train the proposed algorithm on the acquired WCE database from hospital. The WCE database includes 1560 images whose original size is 256×240. In the preprocessing of WCE image, we extended the WCE dataset to 5160 images (including 4260 training images and 900 predict images) via mirror and rotation with MATLAB (R2012a).

We first performed a mirror operation on all the images we collected, and then rotated them clockwise 30 degrees, 45 degrees, 60 degrees, 90 degrees, 120 degrees, etc, and scaled whole experimental WCE image size into 96×96 as input. Output is labels of five diseases including normal, bleeding, ileal erosion, colitis and gastritis. The WCE datasets are grouped as Table II as shown. In order to evaluate the performance of the proposed scheme, we conducted comparative experiments. Table III summarizes the results. The performance is evaluated by mean Average Precision (mAP).

TABLE II. AMOUNT OF TRAIN DATASET AND PREDICT DATASET

Different lesions	Data sets	
	Training data	Predict data
Normal	880	200
Bleeding	770	200
Ileal erosion	860	180
Colitis	920	160
Gastricism	830	160

$$Precision = TP/(TP + FP) \tag{3}$$

where true positive (TP) is the quantity of correct matches between the ground truth and the proposed method. False positive (FP) is the number of images which are in the proposed DCNN network predicted results but not in the ground truth.

TABLE III. MULTI-CLASSIFICATION ACCURACY COMPARISON

Different lesions	Different Algorithms				
	OURs (110MB)	VGG-16 (230MB)	ResNet-50 (270MB)	SVM-LBP	BP-PDFB
Normal	100.00%	100.00%	100.00%	95.50%	96.50%
Bleeding	100.00%	100.00%	100.00%	85.00%	87.70%
Ileal erosion	100.00%	98.65%	100.00%	80.15%	83.35%
Colitis	95.15%	94.50%	100.00%	83.50%	86.25%
Gastricism	96.70%	93.90%	98.75%	84.75%	92.50%
mAP	98.37%	97.41%	**99.75%**	85.78%	89.26%

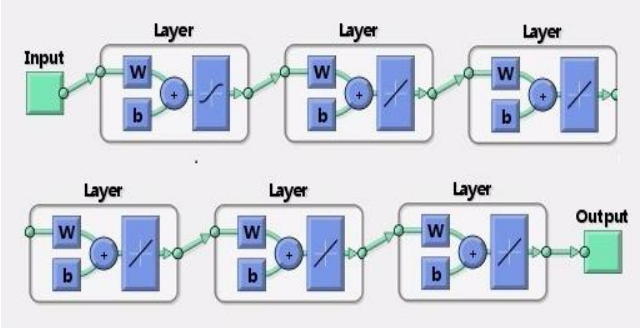

Figure 3. BP neural network's overall architecture map generated by matlab. And the input is a 32-dimensional feature vector extracted by PDFB.

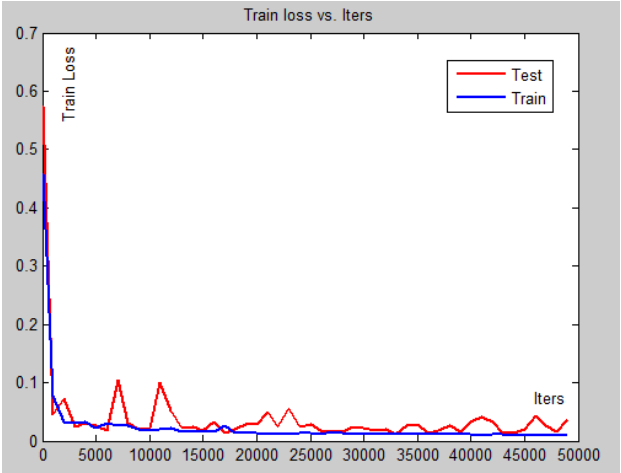

Figure 4. Accuracy curve of training result, the ordinate represents the training accuracy, the abscissa represents the number of iterations. After the iteration exceeds 20,000 times, the recognition rate is stabilized at more than 98.25%, generated by matlab.

SVM is a classical machine learning method to be used for classification task. As a contrast, SVM is employed to recognize five kinds of capsule endoscopy images. The features used in SVM are extracted by LBP. SVM classifier is implemented with a well-known open source library LibSVM [13]. During SVM training process, the model parameters of the penalty parameter c and kernel function g need to be optimized by parameter tuning operation. Given that we know the test set label, so let the two parameters c and g be a range of discrete values. Then, we choose the best test classification parameters to calculate the classification accuracy rate.

Considering BP neural network's weights and thresholds are randomly initialized each time, the result will change a little each time. It is necessary to use the command to save the network weight when finding the ideal result. BP neural network classifier is implemented by MATLAB (R2012a) image processing toolbox. The Nonsubsampled Contourlet transform decomposes images into band-pass directional subband using pyramid directional filter bank (PDFB) and can present edges and texture of images more effectively to extract the hand-crafted feature than wavelet transform. The experimental results are shown in Table III. The recognition accuracy rate is 95.15% by using DCNN designed in this paper. The recognition accuracy rate is 80.15% by using SVM with LBP method and the recognition accuracy rate is 83.35% by using BP Neural Network with contourlet transform method, indicating the correctness of structure network model proposed in this paper.

The BP neural network we use has a six-layer fully connected network. The number of neurons in the first layer or the input layer is seven, and the number of the neurons in the four hidden layers are all 15. The number of neurons in the last layer or the output layer is 5, representing five different types of lesions. The first and second layer's node transfer function is 'logsig'. The third layer's node transfer function is 'tansig' and that of the fourth / fifth / sixth layer is 'logsig'. The weight and bias parameter optimization algorithm is the well-known Levenberg-Marquardt optimization method.

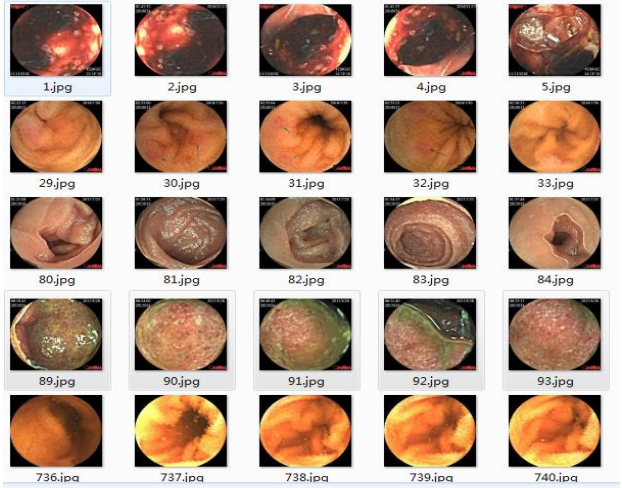

Figure 5. Different lesion types. 1st line: Bleeding WCE image, 2nd line: Ileal erosion WCE image, 3rd line: Gastricism WCE image, 4th line: Colitis WCE image, 5th line: Normal WCE image.

IV. CONCLUSION

In this paper, we present a scheme that can effectively improve the classification accuracy of WCE images contrasting to traditional methods. Our proposed architecture is based on multi-scale feature maps fusion, we try to optimize the training convolution kernel parameters for each layer by quantifying the combined convolutional layers via trainable weight parameters. The experimental results demonstrate that our scheme achieves competitive results compared to state-of-the-art methods in WCE image

classification task. Since the deep learning algorithm requires a large amount of tagged data, more WCE lesion images and advanced network schemes for lesion classification are in great need to further improve the classification accuracy of indistinct lesion images and explore automatic detection methods for lesion localization.

ACKNOWLEDGMENT

This work was partly supported by the National Natural Science Foundation of China (Grant Nos. 61471216 and 61771276), and Special Foundation for the Development of Strategic Emerging Industries of Shenzhen (Grant Nos. JCYJ20170307153940960, JCYJ20170817161845824 and No. JCYJ20150831192224146).

Thanks to Shenzhen Luohu People's Hospital for providing the WCE database.

REFERENCES

[1] H. Jia Sen, Z. Yue Xian, and L. Lei, "An advanced WCE video summary using relation matrix rank," in Biomedical and Health Informatics (BHI), 2012 IEEE-EMBS International Conference on, 2012,pp. 675-678.

[2] Y. Yixuan and M. Q. H. Meng, "Hierarchical key frames extraction for WCE video," in Mechatronics and Automation (ICMA), 2013 IEEE International Conference on, 2013, pp. 225-229.

[3] J. Chen, Y. Wang, and Y. X. Zou, "An adaptive redundant image elimination for Wireless Capsule Endoscopy review based on temporal correlation and color-texture feature similarity," in Digital Signal Processing (DSP), 2015 IEEE International Conference on, 2015, pp. 735-739.

[4] Li S, Kwok J T, Zhu H, et al. Texture classification using the support vector machines[J]. Pattern Recognition, 2003, pp.2883-2893.

[5] Vasconcelos N, Lippman A. A probabilistic architecture for content based image retrieval[C]Computer Vision and Pattern Recognition,2000.Proceedings. IEEE Conference on. IEEE, 2000:216-221 vol.1.

[6] Alex Krizhevsky, Ilya Sutskever, Geoff Hinton. Imagenet Classification with deep convolutional neural networks [J]. in Advances in Neural Information Processing Systems, 2012,pp.1097-1105.

[7] Hariharan B, Arbeláez P, Girshick R, et al. Hypercolumns for Object Segmentation and Fine-grained Localization[J]. In CVPR 2014:447-456.

[8] Zhu R, Zhang R, Xue D. Lesion detection of endoscopy images based on convolutional neural network features[C] International Congress on Image and Signal Processing. IEEE, 2015,pp.372-376.

[9] LeCun Y, Boser B, Denker J S, et al. Back propagation Applied to Handwritten Zip Code Recognition[J]. Neural Computation, 2014, pp.:541-551.

[10] Simonyan K, Zisserman A. Very Deep Convolutional Networks for Large-Scale Image Recognition[J]. Computer Science, 2014.pp.3992-4000.

[11] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In CVPR, 2015,pp.770-778.

[12] Kong T, Yao A, Chen Y, et al. HyperNet: Towards Accurate Region Proposal Generation and Joint Object Detection[C]// Computer Vision and Pattern Recognition. IEEE, 2016:845-853.

[13] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," ACM Trans. Intell. Syst. Technol., vol. 2, pp. 1-27, 2011.

[14] M. D. Zeiler and R. Fergus. Visualizing and understanding convolution neural networks. In ECCV,2014,pp.818-833.

[15] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, et al. Going deeper with convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,2015,pp.1-9.

[16] LeCun Y, Bottou L, Bengio Y, et al. Gradient-Based Learning Applied to Document Recognition,in Proceeding of the IEEE. 1998,pp.2278-2324.

[17] C.Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. In CVPR, 2016,pp.2818-2826.

[18] Bai X, Shi B, Zhang C, et al. Text/non-text image classification in the wild with convolutional neural networks[J]. Pattern Recognition, 2017, pp.437-446.

[19] Hinton GE,Salakhutdinov R.Reducing the dimensionality of data with neural networks.Science,2006,303(7):504-507.

[20] Chen J, Zou Y, Wang Y. Wireless capsule endoscopy video summarization: A learning approach based on Siamese neural network and support vector machine[C]// International Conference on Pattern Recognition. IEEE, 2017:1303-1308.

[21] Velisavljevic V, Dragotti P L, Vetterli M. Directional wavelet transforms and frames[C]// International Conference on Image Processing. 2002. Proceedings. IEEE, 2002:589-592 vol.3.

[22] Do M N, Vetterli M. The contourlet transform: an efficient directional multiresolution image representation[J]. IEEE Transactions on Image Processing A Publication of the IEEE Signal Processing Society, 2005, 14(12):2091-2106.

[23] SUN K,ZHANG S,YAO R,et al.Lesion detection of gastroscopic images based on cost-sensitive boosting[C]//Machine Learning for Signal Processing(MLSP),2011 IEEE International Workshop on.2011:1-6.

[24] Li B, Meng Q H. Computer-based detection of bleeding and ulcer in wireless capsule endoscopy images by chromaticity moments[J].Computers in Biology and Medicine, 2009, 39(2): 141-147.

[25] He K, Zhang X, Ren S, et al. Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2014, 37(9):1904-16.