ORIGINAL RESEARCH



Multi-scale pulmonary nodule classification with deep feature fusion via residual network

Guokai Zhang¹ · Dandan Zhu¹ · Xiao Liu¹ · Mingle Chen² · Laurent Itti³ · Ye Luo¹ · Jianwei Lu^{1,4}

Received: 15 January 2018 / Accepted: 8 November 2018 © Springer-Verlag GmbH Germany, part of Springer Nature 2018

Abstract

The early stage detection of benign and malignant pulmonary nodules plays an important role in clinical diagnosis. The malignancy risk assessment is usually used to guide the doctor in identifying the cancer stage and making follow-up prognosis plan. However, due to the variance of nodules on size, shape, and location, it has been a big challenge to classify the nodules in computer aided diagnosis system. In this paper, we design a novel model based on convolution neural network to achieve automatic pulmonary nodule malignancy classification. By using our model, the multi-scale features are extracted through the multi-convolution process, and the structure of residual blocks allows the network to capture more high-level and semantic information. Moreover, a strategy is proposed to fuse the features from the last avg-pooling layer and the ones from the last residual block to further enhance the performance of our model. Experimental results on the public Lung Image Database Consortium dataset demonstrate that our model can achieve a lung nodule classification accuracy of 87.5% which outperforms state-of-the-art methods.

Keywords Pulmonary nodule classification · Multi-scale · Residual block · Feature fusion

1 Introduction

According to the latest statistical studies (Siegel et al. 2011), the mortality and morbidity of lung cancer have been ranked top worldwide. The 5-year survival rate for lung cancer is only less than 5% at the advanced stage. The primary key to diagnosing the lung cancer at the early stage is to classify the small and spherical lesion structures within the chest cavity which are called the pulmonary nodules. There are two main imaging techniques to detect the pulmonary nodules, one is by using the X-ray imaging and the other one is by

🖂 Ye Luo

yeluo@tongji.edu.cn

- 🖂 Jianwei Lu
- ¹ School of Software Engineering, Tongji University, Shanghai, China
- ² High School Affiliated to Fudan University, Shanghai, China
- ³ Department of Computer Science and Neuroscience Program, University of Southern California, Los Angeles, USA
- ⁴ Institute of Translational Medicine, Tongji University, Shanghai, China

Published online: 12 November 2018

computed tomography (CT). For X-ray imaging, it emits beams through the lung and provides a single view of the lung internal structures, whereas the CT imaging captures three different orientations information by using a rotational scanner which can provide a more detailed and overall structure view of the lung. Specially, CT imaging can facilitate the radiologist to detect the location, size, shape, and internal structures of the pulmonary nodules more easily. Even though the CT imaging techniques have supplied an effective way to detect the pulmonary nodules, the work of labeling the possible malignant nodules is still time-consuming and empirical. In consideration of the huge amounts of CT image slices inspected by the radiologists every day, it could be a continuous pressure for radiologists to handle this task efficiently and accurately.

To alleviate this situation, computer aided detection (CAD) system can be a useful assistant tool to help the radiologists detect and analyze the pulmonary nodules. Many researchers have devoted their efforts to making the procedure efficient and accurate. In this paper, we conduct the research on the automatic malignancy classification of pulmonary nodules which is a crucial and worthy work in clinical risk factors assessment (Gould et al. 2007).

Generally, the methods for pulmonary nodule classification are mainly categorized into two ways: hand-crafted feature classification and automatic feature extraction classification. Compared with the hand-crafted feature extraction, the convolution neural network can effectively capture more linear and nonlinear transformations, and extract more abstract high-level features from the original data (Bengio and Courville 2013). While many convolution neural network models have achieved remarkable performance on pulmonary nodule classification, the main limitation of these work is that pulmonary nodules usually have various sizes, shapes, and locate in different positions as illustrated in Fig. 1. The traditional convolutional neural network (CNN) structures (2017) input the data with a fixed size, which could hamper the model to detect nodule contextual information and scale characteristic. To further improve the medical image classification performance, most of the scale-relevant and discriminative features are extracted by using multi-stream networks (Setio et al. 2016; Dou et al. 2017; Kamnitsas et al. 2017). Multi-stream networks refer to the network with independent parallel sub-networks. Generally, parallel sub-networks usually share the same network structure. The main difference among them is that the inputs of them are different size or different view of images. By this way, the scale-relevant and discriminative features are learned and then fused to output the final classification result. However, those methods need extra pre-training and fine-tuning which may cause more computational complexities and produce more parameters.

In this paper, we exploit an efficient way to learn the multi-scale discriminative features through multiple convolution processes under one neural network stream. Moreover, in order to extract more highly compact and rich information, we fuse the features from the last avg-pooling layer and



Fig. 1 The pulmonary nodule samples with different sizes, shapes, and location variations. A large diameter nodule with hypervascular characteristic is more likely to be a malignant nodule

the ones from the last residual block. The learned features from avg-pooling layer tend to be more global and abstract, meanwhile, high-level features from the last residual block are rich of semantic information. Thus, to add them together is a good fusion way to achieve better model performance. The architecture of the proposed model is illustrated in Fig. 2. The network consists of three components for individual tasks: (1) the multi-scale feature extraction stage consists of four various sizes of convolutions. (2) The highlevel features learning stage consists of a series of residual blocks. (3) The feature fusion stage is to capture global and semantic features.

The contributions of our work can be summarized as follows

- (1) We propose a novel model to classify the nodule malignancy suspicious from CT images. Unlike the previous work with multi-stream networks to extract multiscale features from images, we adopt multi-convolution process to learn the discriminative scale features with fewer computational parameters and complexities.
- (2) We utilize the residual block as a basic unit to extract high-level and semantic information from nodule image.
- (3) A fusion strategy is applied by fusing the features from the last avg-pooling layer and the last residual block to further enhance the performance of our model. The experimental results demonstrate that our model outperforms state-of-the-art methods.

2 Related work

Recently, deep convolution neural network has achieved significant improvement in image classification and object detection task (Krizhevsky and Sutskever 2012). Huang et al. (2017) designed a robust deep learning model for speech emotion recognition. Tang et al. (2018) proposed a novel Canonical Correlation Analysis (CCA)-based multiview convolutional neural network structure for color and depth feature learning. Song et al. (2018) built a very deep network to extract more discriminative features from hyperspectral images by considering the correlated information from different hierarchical layers. Wen et al. (2018) proposed a data-driven vision system based on the deep learning network to recognize objects, the final performance of the proposed system achieved higher recognition accuracy than other methods. The successes of deep convolution neural network in different vision fields also inspire many researchers to exploit the method to analyze the medical images. Song et al. (2015) proposed a multi-scale convolution neural network and graph-partitioning-based method for the segmentation of cervical cytoplasm and nuclei, the



Fig. 2 An overview of the proposed architecture for malignant and benign nodule classification. The input image size is 64×64 px, the $(k@m \times m, s)$ denotes the convolution has k kernels and the convolution size is $m \times m$ with stride s. FC represents the fully connected layer. The network consists of three components for individual tasks:

(1) the multi-scale feature extraction stage consists of four various sizes of convolutions. (2) The high-level features learning stage consists of a series of residual blocks. (3) The feature fusion stage is to capture global and semantic features

experimental results proved that the segmentation method can better deliver promising results than the existing methods. Maninis and Pont-Tuset (2016) adopted a deep retinal image understanding model which was based on the deep convolutional neural networks. By using the network architecture, the model presented super-human performance. Shi et al. (2018) used Multimodal Stacked Deep Polynomial Networks for Alzheimer's Disease diagnosis, experimental results indicated that their network structure was superior to the state-of-the-art multimodel feature learning methods.

For automatic pulmonary nodule classification, many researchers have made great contributions in establishing a more robust and accurate model. The general process of classifying pulmonary nodules usually consists of two steps: extract features from original images and put the extracted features into a classifier to categorize nodules. For example, Uchiyama et al. (2003) used histogram as one important feature to classify the nodules. Way et al. (2006) proposed a pulmonary nodule classification architecture by putting the extracted texture features into a linear discriminant classifier. Messay et al. (2010) analyzed the segmented nodules by combining shape, position, and intensity features, the experimental results achieved the sensitivity score of 82.66% for each average per scan. Han et al. (2015) used texture feature based on a three dimensional image for nodule classification and achieved state-of-the-art classification accuracy. However, those hand-crafted feature extraction methods tend to be subjective and the hyperparameters need to fine-tune by artificial selection.

By contrast with the hand-crafted feature methods, the convolutional neural network can provide an end-to-end training mode to automatically learn the high-level features from the input data. Kumar et al. (2015) applied an autoencoder network to extract the pulmonary nodule features and handled them with a binary decision tree to classify nodules as malignant or benign on LIDC dataset (Armato et al. 2011). The experimental result showed that the designed model achieved a satisfying result at that time. Ciompi et al. (2015) tackling the problem of automatic classification by using 2D pulmonary nodule views, it achieved the performance of AUC (0.868) which was close to human analysis. A novel multi-view convolution neural network (Setio et al. 2016) was proposed and designed to efficiently extract discriminative features from different image orientations. In the fusion stage, the author evaluated the performance on different fusion strategies. The experimental results showed that the late fusion strategy can outperform state-of-the-art methods. Chen et al. (2017)) exploited three multi-task learning (MTL) schemes to leverage multiple features derived from convolution neural network models. Hussein et al. (2017) extracted seven attributes of pulmonary nodules and attained seven scores separately. Shen et al. (2017) investigated the classification of pulmonary nodules by using multi-crop convolution neural network to extract nodule salient information. Experimental results showed that the model can achieve better performance than traditional hand-crafted feature extraction methods.

3 Methodology

3.1 Architecture

In this work, our basic structure of pulmonary nodules classification network is based on residual blocks (He et al. 2016) and the goal of our model is to leverage the network to capture more high-level features hierarchically and semantically. To achieve this goal, we first fuse the multi-convolution outputs to attain rich semantics features at different scales. To be specific, the input nodule image passes through 4 various sizes of convolutions with the receptive field of $(7 \times 7, 14 \times 14, 32 \times 32,$ 64×64) respectively, and each convolution stride is set as 2 pixels during the training process. Compared with the traditional multi-stream network structures, our multi-convolution design can extract diverse scale features with fewer parameters and computational complexities from one single sized nodule image. Then a max-pooling layer is used to reduce the feature maps dimension, it performs with the size of 3×3 , the stride is set as 2 pixels. After the pooling layer operation, eight residual blocks are stacked to further extract the high-level features. The residual blocks have the benefit of training a deeper layer network by using an identity connection, it also allows the network to learn more high-level features from the nodule images. Focusing on the ultimate goal of improving the classification performance in nodules identification, we fully connect outputs of the last residual block and avg-pooling layer to make full use of the global spatial and highly compact features. The detailed network parameters setting is shown in Table 1. The layer of FC_i where $i \in 1, 2$ denotes the operation of flattening the assigned layer and then using the fully connected layer to unify the feature dimension to 128-d. The Merge-layer is used to add the features from the layer FC_1 and FC_2, it also has the same dimension with FC_1 and FC_2 layer. Downsampling is adopted during the Multi-conv, Conv2, Conv3, Conv4 with a stride of 2.

3.2 Teamwork and communication

3.3 Multi-convolution feature extraction strategy

The convolution operation in the neural network is to extract features from the input image with local Table 1 Parameters setting of the network

Layer name	Output size	Filter size	Num
Multi-conv	32 × 32	$\begin{bmatrix} 7 \times 7\\ 14 \times 14\\ 32 \times 32\\ 64 \times 64 \end{bmatrix}$	64
Max-pooling	16 × 16	3×3 , stride 2	_
Conv1	16 × 16	$\begin{bmatrix} 3 \times 3 \\ 3 \times 3 \end{bmatrix} \times 2$	64
Conv2	8 × 8	$\begin{bmatrix} 3 \times 3 \\ 3 \times 3 \end{bmatrix} \times 2$	128
Conv3	4 × 4	$\begin{bmatrix} 3 \times 3 \\ 3 \times 3 \end{bmatrix} \times 2$	256
Conv4	2×2	$\begin{bmatrix} 3 \times 3 \\ 3 \times 3 \end{bmatrix} \times 2$	512
Avg-pooling	1 × 1	3×3 , stride 2	_
FC_1	flatten Conv4, fc, 128-d		
FC_2	flatten Avg-pooling, fc, 128-d		
Merge-layer	Add [FC_1, FC_2], 128-d		
Output-layer	fc+Softmax		

connections and tied weights. Consider the network input is a m * m * c image where the height and width of the image are same sizes with m, and c is the channels' number of the image. If the convolution layer has k kernels and the size of each kernel is p * p * g, where p is not bigger than m, and g has the same size with c. The convolution can produce k feature maps with the size of (m - p)/stride + 1. The operation of convolution can be expressed as

$$y_l^j = max \left(0, b_l^j + \sum_i k_l^{ij} * x_l^i \right), \tag{1}$$

where x_l^i is the *i*-th input feature map, and y_l^j is the *j*-th output feature map. Moreover, k_l^{ij} is the defined kernel, the operator * represents the convolution operation and b_l^j is the bias of y_l^j . Here, *l* denotes the region which weights are shared. In order to increase the nonlinear properties of the decision function, we adopt Relu (Krizhevsky and Sutskever 2012) as the activation function to improve the network performance. The definition of Relu activation function is f(x) = max(0, x), where *x* is the input to the neuron.

The pulmonary nodules often come with various shape structures and complicated texture characteristics which can be a big challenge to classify. To tackle this challenge, we apply multi-convolution process to attain discriminative features from different receptive fields during the first convolution layer operation. The various receptive field of convolution sliding window allows the network to learn more scale-discriminative and semantic features from the

input image data. In fact, if the receptive field of the convolution is too big that may hinder the model to classify the tiny objects. On the contrary, the small receptive field setting can also influence the network to capture global abstract features (Szegedy et al. 2017). Inspired by this theoretical principle, we set the receptive field of four convolutions as $(7 \times 7, 14 \times 14, 32 \times 32, 64 \times 64)$, respectively. The extracted representations from 4 various receptive fields are shown in Fig. 3. The padding is used during the convolution operation to keep the resolution unchanged. From the figure, we can see that the learned features from different receptive fields can capture more abstract and discriminative features. Meanwhile, extracting the features from the first convolution layer could also better preserve the original image location and detailed image information. After four convolutions feature learning, we concatenate the output features from the previous layers and then put them into the max-pooling layer. The concatenate convolution operation is illustrated in Fig. 4. The max-pooling layer is usually used to reduce the dimension of the feature representations, it also has the benefit of avoiding the network being overfitting. The max-pooling layer can be defined as

$$y_{(j,k)}^{i} = \max_{0 \le m, n < s} \left\{ h_{(j \cdot s + m, k \cdot s + n)}^{i} \right\},$$
(2)

where *s* is the pooling region size and the parameter *m* and *n* denote the offset of position $(j \cdot s + m, k \cdot s + n)$. $h^{i}_{(j:s+m,k:s+n)}$



Fig. 4 The proposed multi-convolution structure, after each convolution operation we concatenate the features to capture discriminative scale features

represents the *i*th neuron's input feature map at position $(j \cdot s + m, k \cdot s + n)$, the scalar $y_{j,k}^i$ is the output feature map at position (j, k) in *i*th neuron.

3.4 Residual block representation learning

For the purpose of deepening the network to learn more high-level features with minor network degradation, we apply the residual block to construct a deeper feature learning model. Instead of stacking layers directly, the residual block structure reformulates the layers to learn features by residual functions with reference to the previous layers



Fig. 3 The extracted representations from multi-convolution structure and the four convolutions' size is 7×7 , 14×14 , 32×32 , 64×64 respectively

inputs. Consider H(x) as the underlying mapping of the layers, and x denotes the input of the layer. Let define the residual function as

$$F(x) := H(x) - x \tag{3}$$

Thus the underlying mapping function could be expressed as F(x) + x. Due to directly adding more layers could cause degradation problem, it is likely to be hard to make the identity mapping optimal. While by adopting the residual learning reformulation, the identity mappings can be optimal simply by making the weights of multiple nonlinear layers to zero. This makes it possible for the network to go deeper. The detailed residual learning block is shown in Fig. 5. F + xoperation is achieved by using a shortcut connection. The residual block can not only produce any extra parameter or computational complexity but also improve the model performance remarkably.

3.5 Feature fusion

Avg-pooling layer has been proved effective in giving more global spatial representations and being robust to small transformations of the images (Boureau et al. 2010). In our model, the last avg-pooling layer is to achieve spatial invariance and reduce model computational complexity by averaging the feature vector within a spatial neighborhood. For the last residual block output, it could learn various mid-level or high-level features from different image regions (Huang et al. 2012; Lin and Chen 2013) and contain rich discriminative and semantic information (Szegedy et al. 2015). Inspired by this two advantages, we explicitly fuse these two layers' outputs. In order to fuse those two layers with the same



Fig. 5 The structure of residual block

dimension, we flatten the two output feature maps to one vector and then input them to a fully connected layer respectively. After that operation, we further add them together as one vector to enrich the semantic information. The last fully connected layer is a two-way softmax layer which provides the prediction of probability to be malignant and benign. It can be expressed as

$$y_{i} = \frac{exp(y_{i}')}{\sum_{j=1}^{m} exp(y_{j}')}$$
(4)

where $y'_{j} = \sum_{i=1}^{128} x_{i} \cdot w_{i,j} + b'_{j}$ is the linear combination of input 128 features x_{i} , and y_{i} is the output probabilities. Here, $w_{i,j}$ represents the weight and b'_{j} is a bias, and m = 2. The network loss function is cross-entropy, we optimize the parameters by Stochastic Gradient Descent (SGD).

3.6 Training process

We evaluate our model based on 5-fold cross validation on the selected augmentation data. The data are randomly divided into five same amount of subsets, each of the subsets contains 20% of the entire dataset. During the training process, we use 4 of the total subsets to train our model and leave 1 subset for testing. The original learning rate is 1.0×10^{-3} , it decays by 1.0×10^{-4} over each update. We also set the dropout value (rate = 0.2) to prohibit the model being overfitting. The training process is stopped when the val-loss on the validation dataset does not decrease after 10 epochs. The model is designed based on Tensowflow and the programming language is Python.

4 Experiment

In this section, we design different experiments to validate our model performance. We first test the influence of the different size of data samples, the performance comparisons on three stages' extracted features are also provided to show the effect of different network configurations. Moreover, we test the influence of different depths and classifiers. At last, comparisons with state-of-the-art methods are conducted to validate the effectiveness of our model.

4.1 Dataset

The pulmonary nodule data is from the LIDC dataset which consists of 1010 patients with lung cancer CT scan images. Each of the nodule's detailed information (coordinate, diameter, texture, malignancy etc.) is annotated by four professional radiologists. The nodules' diameter is range from 3mm to 30mm. Due to the variance size of resolution, we use the spline interpolation method to make the spacing with 1 mm \times 1 mm \times 1 mm. In this study, we investigate the malignancy suspiciousness of the nodules. The malignancy score of each nodule is annotated by radiologists which are rated from 1 to 5. We adopt the voting strategy as the final decision result. If more than two radiologists annotated the nodule score over 3, we regard the nodule as malignant. On the contrary, the nodule is regarded as benign. In total, there are 195 malignant nodules and 158 benign nodules respectively. We discard the nodules which have the same votes. In order to reduce the computational complexity, we extract the central transection of each nodule voxel, and the nodule image size is 64×64 .

We use data augmentation to generate more data to avoid the network being overfitting. The data augmentation we used is same with previous work (Xu et al. 2017). Specifically, we randomly flip and magnify the image by zooming 0.2. The rotation is θ , and $\theta \in [30^\circ, 60^\circ]$. The translation is γ which ranges from -6 to 6 voxels. At last, the number of malignant nodules is 390, the number of benign nodules is 316.

4.2 Evaluation metrics

The model performance is measured by four evaluation metrics during the 5-fold cross validation process. Let define TP, FP as true positives and false positives respectively. The FN, TN denotes as false negatives and true negatives separately.

(1) Classification accuracy is the ratio of the correct sample numbers divided by the total sample numbers. It can be described as

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$
(5)

(2) Sensitivity is a score which measures the classifier predicts true positive correctly as the true positive.

$$Sensitivity = \frac{TP}{TP + FN}$$
(6)

(2) Specificity is the score correctly predicted as true negative out of all negative samples.

$$Specificity = \frac{TN}{TN + FP}$$
(7)

(4) Receiver operating characteristic curve (ROC) is a graphical plot which represents the ability of the classifier. AUC score is the area under the ROC curve, it's also an important evaluation metric to estimate the performance of the model.

4.3 Analysis on different size of samples

In deep learning task, it is a crucial issue for the size of training data samples. We first evaluate the model performance on different sizes of sample data. The detailed setting of sample data is divided into three sub-datasets: entire dataset, twice dataset and triple size of dataset. The size of the original dataset is 353. We use data augmentation methods to augment the data to 706 and 1059, respectively. The final comparison of different samples is shown in Table 2. The result shows that adding more data could improve the performance to a certain extent but not always improve. In this paper, we use the 706 data samples as the experimental data.

4.4 Comparison with multi-convolution process

The receptive field of convolution plays an important role in learning multi-scale features. Compared with the traditional multi-stream convolution neural networks, we simply use multi-convolution with different sizes to capture the discriminative scale features. In this section, we investigate how the convolution numbers affect the final performance of the model. In our model, four different size of convolutions $\{7 \times 7, 4 \times 14, 32 \times 32, 64 \times 64\}$ are adopted in the first convolution layer. We believe that the extracted features from the first layer contain much more location information of the original nodule image, and that information can be a crucial point in the nodule classification task. Figure 6 illustrates the results of different convolutions' configurations, the first column of the graph is ResNet-18 which has the same structure as previous work (Nibali et al. 2017). The second column (No convolution fusion, $\{7 \times 7\}$) is the designed model without multi-convolution design. The rest columns of the graph are shown the results with two ($\{7 \times 7, 14 \times 14\}$ }), three $(\{7 \times 7, 14 \times 14, 32 \times 32\})$, four $(\{7 \times 7, 14 \times 14, 32 \times 32\})$ 32×32 , 64×64) and five ({7 × 7, 14 × 14, 23 × 23, $32 \times 32, 64 \times 64$) convolutions fusion respectively. The result demonstrates that different numbers of convolutions could be an important factor in improving the classification performance of the model. Table 3 compares the final AUC score with different fusion configurations. Although the AUC score is not significantly improved by fusion setting, the highest accuracy is achieved by four convolutions fusion setting, it also has the best sensitivity score which is

Table 2 Comparisons with various data sizes

Datasize	353	706	1059
Accuracy (%)	84.1	87.5	86.3
Sensitivity (%)	83.2	94.1	85.2
Specificity (%)	74.2	81.7	83.3
AUC	0.88	0.91	0.90



Fig. 6 Comparisons with accuracy, sensitivity, specificity. **a** ResNet-18; **b** no convolution fusion; **c** two convolutions fusion; **d** three convolutions fusion; **e** four convolutions fusion; **e** five convolutions fusion

 Table 3
 Comparispons with different convolution fusion configurations

Method	AUC
ResNet-18	0.898
No convolution fusion	0.883
Two convolutions fusion	0.886
Three convolutions fusion	0.887
Four convolutions fusion	0.901
Five convolutions fusion	0.897

a critical target in clinical diagnosis. The five convolutions fusion strategy seems to achieve a comparable result with four convolutions fusion strategy. However, blindly adding more different size of convolutions could increase the complexity of the network. In this work, we use the four convolutions fusion strategy as the basic structure.

4.5 Evaluation on different depth of network

The depth of representations is of crucial importance to image classification. Recent research (Simonyan and Zisserman 2014; Szegedy et al. 2015) proves that a deeper network is more liable to achieve better performance. Inspired by the importance of depth, we conduct corresponding experiments to explore the relation between network depth and nodule classification performance. In this paper, we increase the network depth by adding more residual blocks. The experimental result is illustrated in Table 4. It is interesting to find that the classification performance is not always improved as the depth increases. It could be explained that for the specific image classification task, directly increase the network depth could only gain better performance at a certain point but could not improve all the time. Another possible reason
 Table 4
 Proposed model classification performance with different numbers of residual blocks

Method	Residual Blocks-4	Residual Blocks-8	Residual Blocks-16
Accuracy (%)	82.0	87.5	84.6
Sensitivity (%)	85.3	94.1	87.1
Specificity (%)	78.9	81.7	82.2
AUC	0.889	0.901	0.890

 Table 5
 Evaluation on different fusion strategies

Method	Accuracy (%)
No fusion	83.1
Avg-pooling (add)	87.5
Avg-pooling (concatenate)	82.6
Avg-pooling (average)	84.2
Max-pooling (add)	83.3
Max-pooling (concatenate)	84.1
Max-pooling (average)	83.7

could be explained is that the training data size is limited, and blindly deepen the network may cause overfitting problem. We would try to collect more pulmonary nodule data in the future to validate this hypothesis.

4.6 The effectiveness of feature fusion

Pooling layer which has been an important part in convolution neural network structure can yield useful abstract features for object detection and image classification. Generally, there are usually two types of pooling layers in convolution neural network, the average pooling layer and the max pooling layer. Avg-Pooling layer can efficiently attain invariance of image transformations and extract more compact representations. Max pooling layer can be more robust to image scale changes. Meanwhile, with the deepening of the network level, the semantic and high-level features are back propagated which also can be crucial to image classification. Motivated by these two conditions, in our model we fuse the last pooling layer and residual block output features to further improve the nodule classification ability. To verify the effectiveness of our fusion strategy, we design different fusion configurations to conduct the experiment. The comparison results are shown in Table 5. From the result, we find that fuse the output features from pooling layer and last residual block could improve the classification accuracy. Different fusion strategies tend to have various impact on the final classification performance. The best experimental result is fusing the average pooling layer and last residual



Fig. 7 ROC comparisons with different fusion strategies

block by adding manner. It has achieved 87.5% accuracy which is higher than other fusion manners.

We further evaluate the performance of various average pooling layer configurations based on the ROC curve metric. The detailed result is shown in Fig. 7. From the result, we can see that the adding fusion strategy could achieve a minor AUC score improvement compared with other fusion configurations. That could be explained that adding features from pooling layer and last residual block output could enrich more detailed and hierarchical features to the model and finally enhance the classification performance.

4.7 Performance with different classifiers

The neural network has been proved a powerful tool for image feature extraction. Generally, most of the neural network structure uses the softmax or sigmoid as the final activation to output the classification probabilities. In this section, we perform extra experiments to investigate the influence of various classifiers on the final classification accuracy. We adopt another two classifiers support vector machine (SVM) and random forest (RF) to conduct the experiment. For SVM classifier, we use the rbf as the kernel function. The parameter C is optimized by grid search and the value is in the range $[2^{-10}, 2^{-5} \dots 2^5, 2^{10}]$. We also evaluate the influence of output feature dimensions on the final classification performance. We adopt 32, 64, 128, 256 output features as the experimental settings. The detailed results are shown in Table 6. It is obvious that the 128 output features with SVM classifier achieve the best performance with 88.2% accuracy. This is probably because the SVM classifier could efficiently classify the output features with multiple kernel trick, and implicitly map the features into high dimensional feature spaces. It is also noteworthy that the different dimensions of output features could be an important setting parameter to the model performance. Overall, the 128 output

 Table 6
 Performance
 on
 different
 classifiers
 and
 feature
 output

 dimensions

Method	Dim.	Accuracy (%)
FC+SVM	32	84.1
	64	84.3
	128	88.2
	256	84.2
FC + RF	32	83.1
	64	84.0
	128	87.2
	256	84.5
FC + Softmax	32	83.1
	64	84.3
	128	87.5
	256	84.8

features perform better than other parameter settings with the same classifier.

4.8 Comparison with state-of-the-art methods

In order to benchmark the performance of our model to the related methods, we compare our model with some basic neural network structures. The first comparison method is the traditional convolution neural network which has the same structure as previous work (Yang et al. 2016). In this method, four groups of convolution and max-pooling layers extract the features hierarchically from the input image, and subsequently input the extracted features into a fully connected layer with a softmax activation function to classify the nodules. The second method is ResNet-18 and related work (Nibali et al. 2017) on nodule classification by this method has proven its effectiveness. We also adopt Vgg16 (Simonyan and Zisserman 2014) which has achieved great success in the computer vision field to conduct our experiment. The ROC comparison result is shown in Fig. 8. From the result, we observe that our designed model gain a little higher AUC score than other classic network structures.

We also evaluate the performance of these models based on other metrics respectively. The detailed results are presented in Fig. 9. By using our method, the classification accuracy and sensitivity could be enhanced to 87.5% and 94.1% respectively which achieve the best performance of these models. It is noteworthy that the Vgg16 structure's performance is not as well as expected, the reason could be the initial Vgg16 network input size is 224×224 . During the training, we previously resize the nodule images from 64×64 to 224×224 which could hamper the network to extract more global information from the original nodule images.



Fig. 8 ROC comparisons with traditional neural network structures

Further comparisons are conducted with other state-ofthe-art methods on the same dataset. The detailed comparison results are given in Table 7. The result demonstrates that the proposed model could gain better accuracy and sensitivity performance than the other methods. While our AUC and specificity are lower than some of the other methods, this is mainly because these methods adopt 3D convolution neural network to classify the nodules. We believe that our model could obtain same or even better performance when we change our structure to 3D convolution neural network. Figure 10 shows the predict suspiciousness of some nodule samples. The higher score they get, the more malignancy suspicious they would be.

5 Conclusion

Pulmonary nodule malignancy classification is a challenging task due to its different shapes, sizes, and location variations. In this paper, we propose a novel nodule malignancy classification model based on the convolution neural network. We demonstrate that the multi-convolution process and fusion feature strategy could improve the performance of nodule classification, the detailed experimental results have proved the effectiveness of our model. Compared with other state-of-the-art methods that are evaluated on the opening LIDC dataset, our proposed model



Fig. 9 Compared accuracy, sensitivity, specificity with traditional structures. a CNN; b ResNet; c Vgg-16; d Our proposed model



Fig. 10 Results of pulmonary nodules malignancy prediction and the p-value below each image means the probabilityp of being malignant

could gain comparable or even pbetter performance. The further investigations will be focused on other 3D deep learning networks and collecting more clinical pulmonary nodule data to obtain more promising classification performance.

Table 7Performancecomparisons with differentnodule classification methodson the LIDC dataset

Method	Accuracy (%)	Sensitivity (%)	Specificity (%)	AUC
(Kumar et al. 2015)	75.0	83.3	_	_
(Han et al. 2015)	-	89.4	86.0	0.941
(Dhara et al. 2016)	-	89.7	86.3	0.951
(Chen et al. 2017)	86.8	60.3	95.4	-
(Shen et al. 2017)	87.1	77.0	93.0	0.930
Proposed model	87.5	94.1	81.7	0.901

Acknowledgements This work has been supported by the General Program of National Natural Science Foundation of China (NSFC) under Grant nos. 61572362, 81571347, and 61806147, the Central Universities under Grant no. 22120180012.

References

- Armato SG III, McLennan G, Bidaut L et al (2011) The lung image database consortium (LIDC) and image database resource initiative (IDRI): a completed reference database of lung nodules on CT scans. Med Phys 38(2):915–931. https://doi.org/10.1118/1.35282 04
- Bengio Y, Courville A et al (2013) Representation learning: a review and new perspectives. IEEE Trans Pattern Anal Mach Intell 35(8):1798–1828. https://doi.org/10.1109/TPAMI.2013.50
- Boureau YL, Ponce J, LeCun Y (2010) A theoretical analysis of feature pooling in visual recognition. In: Proceedings of the 27th international conference on machine learning, pp 111–118
- Chen S, Qin J, Ji X et al (2017) Automatic scoring of multiple semantic attributes with multi-task feature leverage: a study on pulmonary nodules in CT images. IEEE Trans Med Imaging 36(3):802–814. https://doi.org/10.1109/TMI.2016.2629462
- Ciompi F, de Hoop B, van Riel et al (2015) Automatic classification of pulmonary peri-fissural nodules in computed tomography using an ensemble of 2D views and a convolutional neural network out-of-the-box. Med Image Anal 26(1):195–202. https://doi. org/10.1016/j.media.2015.08.001
- Dhara AK, Mukhopadhyay S, Dutta A, Garg M, Khandelwal N (2016) A combination of shape and texture features for classification of pulmonary nodules in lung CT images. J Digit Imaging 29(4):466–475. https://doi.org/10.1007/s10278-015-9857-6
- Dou Q, Chen H, Yu L, Qin J, Heng PA (2017) Multilevel contextual 3-D CNNS for false positive reduction in pulmonary nodule detection. IEEE Trans Biomed Eng 64(7):1558–1567. https://doi. org/10.1109/TBME.2016.2613502
- Gould MK, Ananth L, Barnett PG (2007) A clinical model to estimate the pretest probability of lung cancer in patients with solitary pulmonary nodules. Chest 131(2):383–388. https://doi.org/10.1378/ chest.06-1261
- Han F, Wang H, Zhang G et al (2015) Texture feature analysis for computer-aided diagnosis on pulmonary nodules. J Digit Imaging 28(1):99–115. https://doi.org/10.1007/s10278-014-9718-8
- He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 770–778
- Huang GB, Lee H, Learned-Miller E (2012) Learning hierarchical representations for face verification with convolutional deep belief networks. In: Computer vision and pattern recognition (CVPR), 2012 IEEE conference on IEEE, pp 2518–2525. https://doi.org/10.1109/CVPR.2012.6247968
- Huang Y, Tian K, Wu A, Zhang G (2017) Feature fusion methods research based on deep belief networks for speech emotion recognition under noise condition. J Ambient Intell Humaniz Comput. https://doi.org/10.1007/s12652-017-0644-8
- Hussein S, Cao K, Song Q, Bagci U (2017) Risk stratification of lung nodules using 3D CNN-based multi-task learning. In: International conference on information processing in medical imaging. Springer, Berlin, pp 249–260. https://doi.org/10.1007/978-3-319-59050-9_20
- Kamnitsas K, Ledig C, Newcombe VF et al (2017) Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation. Med Image Anal 36:61–78. https://doi.org/10.1016/j. media.2016.10.004

- Krizhevsky A, Sutskever I et al (2012) Imagenet classification with deep convolutional neural networks. Adv Neural Inf Process Syst 2015:1097–1105
- Kumar D, Wong A, Clausi D et al. (2015) Lung nodule classification using deep features in CT images. In: Computer and robot vision (CRV), 2015 12th conference on, IEEE, pp 133–138. https://doi. org/10.1109/CRV.2015.25
- Lin M, Chen Q et al (2013) Network in network. arXiv:1312.4400 (arXiv preprint)
- Maninis KK, Pont-Tuset J et al. (2016) Deep retinal image understanding. In: International conference on medical imagecomputing and computer-assisted intervention. Springer, pp 140–148.https://doi. org/10.1007/978-3-319-46723-8_17
- Messay T, Hardie RC, Rogers SK (2010) A new computationally efficient CAD system for pulmonary nodule detection in CT imagery. Med Image Anal 14(3):390–406. https://doi.org/10.1016/j.media .2010.02.004
- Nibali A, He Z, Wollersheim D (2017) Pulmonary nodule classification with deep residual networks. Int J Comput Assist Radiol Surg 12(10):1799–1808. https://doi.org/10.1007/s11548-017-1605-6p
- Setio A, Ciompi F, Litjens G et al (2016) Pulmonary nodule detection in CT images: false positive reduction using multi-view convolutional networks. IEEE Trans Med Imaging 35(5):1160–1169. https ://doi.org/10.1109/TMI.2016.2536809
- Shen W, Zhou M, Yang F et al (2017) Multi-crop convolutional neural networks for lung nodule malignancy suspiciousness classification. Pattern Recognit 61:663–673. https://doi.org/10.1016/j.patco g.2016.05.029
- Shi J, Zheng X, Li Y, Zhang Q, Ying S (2018) Multimodal neuroimaging feature learning with multimodal stacked deep polynomial networks for diagnosis of Alzheimer's disease. IEEE J Biomed Health Inform 22(1):173–183. https://doi.org/10.1109/ JBHI.2017.2655720
- Siegel R, Ward E, Brawley O et al (2011) Cancer statistics, 2011: the impact of eliminating socioeconomic and racial disparities on premature cancer deaths. CA Cancer J Clin 61(4):212–236. https:// doi.org/10.3322/caac.20121
- Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556 (arXiv preprint)
- Song Y, Zhang L, Chen S, Ni D, Lei B, Wang T (2015) Accurate segmentation of cervical cytoplasm and nuclei based on multiscale convolutional network and graph partitioning. IEEE Trans Biomed Eng 62(10):2421–2433. https://doi.org/10.1109/ TBME.2015.2430895
- Song W, Li S, Fang L, Lu T (2018) Hyperspectral image classification with deep feature fusion network. IEEE Trans Geosci Remote Sens 56(6):3173–3184. https://doi.org/10.1109/TGRS.2018.2794326
- Szegedy C, Ioffe S, Vanhoucke V, Alemi AA (2017) Inception-v4, inception-resnet and the impact of residual connections on learning. AAAI 4:12
- Szegedy C, Liu W, Jia Y et al (2015) Going deeper with convolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1–9
- Tang L, Yang ZX, Jia K (2018) Canonical correlation analysis regularization: an effective deep multi-view learning baseline for RGB-D object recognition. IEEE Trans Auton Ment Dev. https:// doi.org/10.1109/TCDS.2018.2866587
- Uchiyama Y, Katsuragawa S, Abe H, Shiraishi J et al (2003) Quantitative computerized analysis of diffuse lung disease in high resolution computed tomography. Med Pyhs 30(9):2440–2454. https:// doi.org/10.1118/1.1597431
- Way TW, Hadjiiski LM et al (2006) Computer-aided diagnosis of pulmonary nodules on CT scans: segmentation and classification using 3D active contours. Med Pyhs 33(7Part1):2323–2337. https ://doi.org/10.1118/1.2207129

Wen Z, Liu D, Liu X, Zhong L, Lv Y, Jia Y (2018) Deep learning based smart radar vision system for object recognition. J Ambient Intell Humaniz Comput. https://doi.org/10.1007/s12652-018-0853-9

Xu Y, Zhang G, Li Y, Luo Y, Lu J (2017) A Hybrid Model: DGnet-SVM for the Classification of Pulmonary Nodules. In: International conference on neural information processing. Springer, pp 732–741. https://doi.org/10.1007/978-3-319-70093-9_78

Yang H, Yu H, Wang G (2016) Deep learning for the classification of lung nodules. arXiv:1611.06651 (arXiv preprint)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.