Neurocomputing 197 (2016) 221–231

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

A deep feature based framework for breast masses classification

Zhicheng Jiao, Xinbo Gao *, Ying Wang, Jie Li

Lab of Video and Image Processing Systems, School of Electronic Engineering, Xidian University, Xi'an, China

ARTICLE INFO

Article history:
Received 26 November 2015
Received in revised form
29 January 2016
Accepted 9 February 2016
Available online 17 March 2016

Keywords:
Deep learning
Convolutional neural network
Breast mass classification
Computer-aided diagnosis
Feature visualization

ABSTRACT

Characteristic classification of mass plays a role of vital importance in diagnosis of breast cancer. The existing computer aided diagnosis (CAD) methods used to benefit a lot from low-level or middle-level features which are not that good at the simulation of real diagnostic processes, adding difficulties in improving the classification performance. In this paper, we design a deep feature based framework for breast mass classification task. It mainly contains a convolutional neural network (CNN) and a decision mechanism. Combining intensity information and deep features automatically extracted by the trained CNN from the original image, our proposed method could better simulate the diagnostic procedure operated by doctors and achieved state-of-art performance. In this framework, doctors' global and local impressions left by mass images were represented by deep features extracted from two different layers called high-level and middle-level features. Meanwhile, the original images were regarded as detailed descriptions of the breast mass. Then, classifiers based on features above were used in combination to predict classes of test images. And outcomes of classifiers based on different features were analyzed jointly to determine the types of test images. With the help of two kinds of feature visualization methods, deep features extracted from different layers illustrate effective in classification performance and diagnosis simulation. In addition, our method was applied to DDSM dataset and achieved high accuracy under two objective evaluation measures.

1. Introduction

1.1. Breast mass classification

Being the second cause of death, breast cancer is one of the most common cancers in women. According to a world health organization (WHO) report, breast cancer accounts for 22.9% of diagnosed cancers and 13.7% of cancer related death worldwide [1]. To improve the five-year and ten-year survival rate and to relieve great suffering of patients, the early diagnosis is of crucial importance. Being a process of utilizing low-energy X-rays to examine the human breast, mammography is the most widely used screening and diagnostic tool in both clinical and scientific fields. In order to analyze such an amount of mammograms generated daily in medical centers and hospitals, traditional solution for this challenge is that radiologists have to browse all these images day and night. The next several diagnosis processes also exhaust the physicians, causing the diagnosis to be highly susceptible to errors. This situation also troubles physicians in other fields, and computer-aided diagnosis (CAD) systems have been playing more and more important parts in assisting and improving physicians' work. In previous works, Doi [2] considered that CAD had become one of the major research subjects in medical imaging and diagnostic radiology. Ginneken et al. [3] pointed out that CAD systems were of great help in diagnosis of chest radiography. Jiang et al. [4] and Chan et al. [5] obtained the conclusion that CAD could be used to improve radiologists' performance in breast cancer diagnosis. Identifying benign and malignant masses is among the core contents in diagnosis using mammography. Meanwhile, the building of systems which can effectively assist to do mass classification is one of the hotspots in the mammography related CAD field. Therefore, designing better classification algorithms and frameworks has been attracting more and more attention.

However, as shown in Fig. 1, owing to the diversity in appearance, it is difficult to distinguish the malignant masses from benign ones. A number of researchers and their research teams have been devoted to designing learning and classifying framework to overcome this difficulty. Rangayyan et al. [6] proposed using morphological features to characterize the roughness of tumor boundaries and applied them in classification tasks; Mavroforakis et al. [7] used linear, neural and SVM classifiers to classify masses with the help of textural features and conducted fractal dimension analysis; Timp et al. [8] came up with a novel method exploring the temporal change features among mammography series by regional registration; Rojas-Domínguez et al. [9] performed the analysis of the gradient orientation, fuzziness,

* Corresponding author.
E-mail address: xbgao@mail.xidian.edu.cn (X. Gao).

http://dx.doi.org/10.1016/j.neucom.2016.02.060
0925-2312/

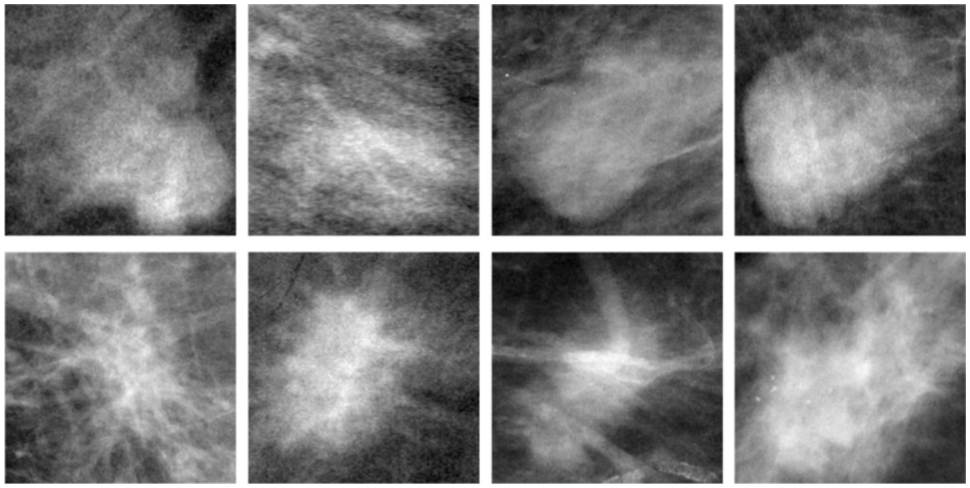


Fig. 1. Examples of benign and malignant breast mass images. Instances in the first row are benign masses while ones in the second row are malignant.

speculation, and mutual information of mass margins; Employing BI-RADS mammographic features with SVM-REF classifier, Yoon et al. [10] achieved a good performance in DDSM; Ramirez-Villegas et al. [11] chose SVM and neural-based classification methods combining Wavelet packet energy, Tsallis entropy and statistical parameterization feature analysis; benefitting more from data structure high accuracy was also reported by Wang et al. [12] in the way of formulating this task into one second order cone programming problem. Verma et al. [13,14] extracted various kinds of features such as density, morphology, abnormity assessment rank, and so on, for description of the masses. Then the soft neural network and soft clustered based direct learning method were employed to do the classification. Wang et al. [15] proposed a latent feature mining based method which characterized spatial and marginal information effectively and achieved good results. More recently, Beura et al. [16] proposed a scheme utilizing 2D-DWT and GLCM in succession to derive feature matrix form mammograms for further classification. Besides, Xie et al. [17] applied extreme learning machine method to improve the performance of mass classification tasks.

Methods which were mentioned above used to extract and utilize low-level features such as margin, texture and so on, or middle-level features such as shape and some variants of bag of words (BOW) [18–20], which has been proven to be effective by Avni et al. [21]. Then the features are introduced into different kinds of classifiers to categorize masses. Although the wide range of traditional handcraft features seem like building a good description of an image, there has been a significant gap existing between these features and cognitive behaviors of physicians. And they do not seem to cover the basic strategies [2] for development of CAD methods and techniques. Strategies aiming at achieving detection and quantitation of lesions in medical images should be based on the understanding of image readings by radiologists. In the real diagnosis process, doctors usually glance at the X-ray first to get preliminary understanding of it. Then several regions that might contain lesions would attract more attention, and the overall look of these regions and details of the entire image would leave impression on and result in different levels of knowledge in the physician's brain. To make the judgement of whether a mass is benign or malignant, doctors used to combine the varying levels of knowledge and awareness with previous experience in similar tasks. The procedure above is similar to that described by the attending doctor we consult from, and it agrees with two representational diagnosis methods: symptom comparisons and anti-diastole. Among these processes, the hierarchical impression of mass images and information processing of human brain are difficult to be specified with traditional features and related methods. Nevertheless, all these things which could only be unspeakable account for a lot in the real diagnosis. So, methods utilizing hierarchical representations and a similar decision mechanism to the real diagnosis may be better choices.

In addition, various types of traditional features can improve classification performance in most situations, but they may have some negative impacts owing to the incompatibility. For example, incompatible extraction methods and corresponding features are less explicable when they are combined directly in a unified framework. However, designing an effective feature fusion strategy is also exhausting in previous papers. On the contrary, hierarchical frameworks could put features extracted from different levels together to form a more explainable and unified structure, avoiding fusing features directly with different models.

1.2. Deep learning and deep learning on biomedical image

From the year of 2006 on, a new machine learning paradigm, named deep learning [22–24], has been playing a much more important role in the academic community. And it has become a huge tide of technology trend in the field of big data and artificial intelligence. Simulating the hierarchical structure of human brain and its data processing mechanism which transfers information from lower level to higher level, deep learning introduces more semantic information to the final representations. Thus deep structure makes significant breakthroughs on image understanding, speech recognition, natural language processing and many other areas [22]. The fever of deep learning has been sweeping the world and has been attracting more attention of top researchers. As results of all these efforts, a few outstanding deep structures are proposed and prove to be successful, such as convolutional neural network (CNN) [25], sparse autoencoder (SAE) [26], restricted boltzmann machine (RBM) [27,28], and so on.

As aforementioned, a convolutional neural network (CNN) is a representative structure of deep models. It is a type of feedforward artificial neural network where individual neurons are tiled in a way that they respond to overlapping regions called receptive fields [47] in the visual field. And it is inspired by biological processes and is a variation of multilayer perceptrons which are designed to reduce preprocessing.

Having been developed for more than three decades, CNN has become an outstanding method. The powerful structures were introduced by Fukushima [29]. CNN was later improved by LeCun et al. [30]. The famous leNet-5 [31] being in form of CNN obtained huge success in recognizing checking numbers. However, given

more complex tasks, the computational complexity of network would continue to increase as network become deeper, causing the main limitation of deep structure at that time. In the subsequent years, development of computing power and optimization methods makes it possible to train a deeper CNN which is powerful enough [32] to fulfill other more complicated tasks. Specially, CNN has been applied in aspects of visual object recognition and image classification tasks and has achieved superior performance [33–35]. When it comes to the fields of biomedical image processing, many breakthroughs are also made by the powerful structure. Jain et al. [36], Jain and Seung [37], and Helmstaedter et al. [38] applied CNNs to restore and segment the volumetric electron microscopy images. In the next few years, CNN based on patches was also applied by Ciresan et al. [39,40] to detect mitosis in breast histology images. More recently, Zhang et al. [41] proposed a CNN based method aiming at segmenting infant brain tissue images in the isointense stage. A patch based CNN was trained and applied to classify each pixel in the image to finish segmentation. And results show that their proposed model significantly outperformed previous methods on infant brain tissue segmentation.

Being in much the same way of brain's information processing and cognitive mechanism, deep learning could provide more effective features for computer vision tasks, such as detection and classification. And deep neural networks have been proved to be more similar to the primate visual system and hierarchical sensory processing systems in brain [42,43]. Inspired by the achievements of CNN in other fields and its similarity to brain, we proposed a deep feature based framework for breast masses classification. In the proposed scheme, a convolutional neural network (CNN) was trained on a large number of natural images and was fine-tuned on a subset of breast mass images. The training strategy was chosen to overcome shortage of breast images. In this framework the data augmentation [60] operation was also introduced and played a role. Then features of masses were extracted from different hierarchical levels of this model, with the help of which two classifiers were trained for the decision procedure. And we applied a strategy in the decision mechanism, which fused the outcomes from different classifiers to finish the classification.

2. Network training and decision mechanism

Our proposed breast mass classification method consists of a hierarchical representation network and a series of decision mechanism for these features. The fundamental blocks in this network are introduced respectively. Then, training for our CNN and decision mechanism is detailed as follows.

2.1. Network training.

As shown in Fig. 2, the CNN architecture we used in this paper mainly includes 3 kinds of operational units: convolution, pooling and Rectified Linear Unit (ReLU) [44] activation. Each convolution layer in a CNN contains some of these units. Being the indispensable component in CNN frameworks, convolution blocks simulate orientation-selective simple cells in primary visual cortex [25]. It computes the convolution of the input map x with a bank of K multi-dimensional filters f and biases b. Here, H, W, D respectively stand for the height, width and depth of input map x, while H', W', D stand for the height, width and depth of convolution filters, d, d', d'' stand for the channel index of filters, input map and output map. Besides, H'', W'', K stand for the scale of output map y of this layer. $x \in R^{H\times W\times D}$, $f \in R^{H'\times W'\times D\times K}$, $y \in R^{H''\times W''\times D\times K}$, $W'' = W - W' + 1$, $H'' = H - H' + 1$:

$$y_{i''j''d''} = b_{d''} + \sum_{i'=1}^{H'} \sum_{j'=1}^{W'} \sum_{d'=1}^{D} f_{i'j'd} \times x_{i''+i'-1,j''+j'-1,d',d''} \quad (1)$$

Pooling is also an important kind of operation in CNN structure. The pooling unit whose role is similar to complex cells in brain visual cortex has a number of variations [25]. And all these versions have been discussed and compared in previous papers. From

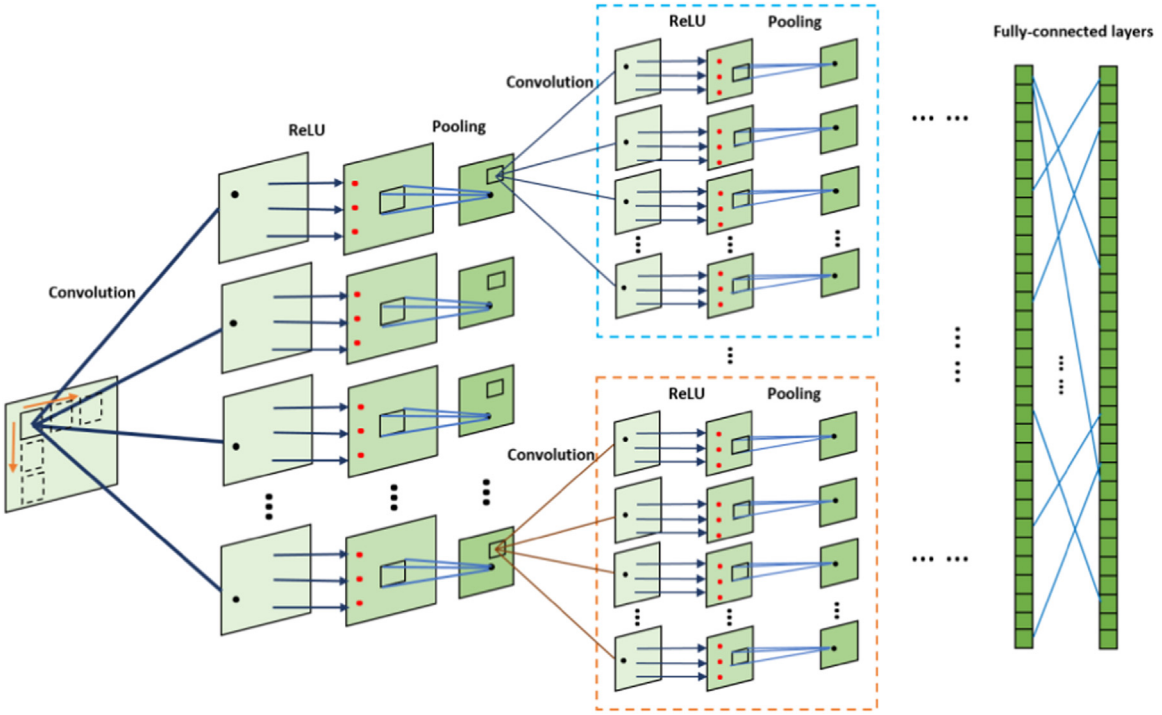


Fig. 2. The main components and connections between them in a deep CNN model. (Convolution) stands for the convolution operation in each layer. (ReLU) is the activation function we chose. (Pooling) is the max pooling operation. These operation units connected successively in each layer. In addition, (Fully-connected layers) which are similar to the traditional neural networks appeared in the last few layers.

Table 1
Detailed parameters of each layer.

Name	Filter size	Filter dimension	Stride	Padding
Conv1	7	1	2	0
ReLU1	1		1	0
Pooling1	3		3	2
Conv2	5	96	1	2
ReLU2	1		1	0
Pooling2	2		2	1
Conv3	3	256	1	1
ReLU3	1		1	0
Conv4	3	384	1	1
ReLU4	1		1	0
Conv5	3	384	1	1
ReLU5	1		1	0
Pooling5	3		3	1
Fc6	1	256	1	0
ReLU6	1		1	0
Fc7	1	2048	1	0
ReLU7	1		1	0
Fc8	1	2	1	0

all these variations, we chose the one named max pooling which has been widely applied in many successful models. It computes the maximum response of each feature channel in x in a $H' \times W'$ patch. Here, H, W, D and H'', W'', K represent the scale of input and out map of this kind of operation. $x \in R^{H \times W \times D}$, $y \in R^{H'' \times W'' \times K}$:

$$y_{i''j''d} = \max_{H' \times W'} x_{i''+i'-1, j''+j'-1, d} \tag{2}$$

Besides, the pointwise activation function is another fundamental component in the model. It simulates excitability of neurons in brain when excited by stimuli. There are also a number kinds of activation functions in deep models. We chose the one named Rectified Linear Unit (ReLU) [44] which has been proved more efficient in computation in the papers of [45] and [46]. Here, y_{ijd} stands for the response in output map of this layer with input x_{ijd} in the corresponding location.

$$y_{ijd} = \max\{0, x_{ijd}\} \tag{3}$$

The receptive fields [47] of different neurons in the network appear in the forms of convolutional and pooling kernels which differ in both size and weight in different layers.

Beyond the aforementioned components, there are several other kinds of layers in this model. With the help of dropout operation [33,46,49], we could reduce overfitting of the network and learn more robust features. Besides, a cross-channel normalization operator is also implemented at each spatial location across all feature maps of the same layer to gain a better description of input. The last layers are fully-connected ones which were followed by logarithmic loss to be minimized.

Our feature representation net is a CNN inspired by [33,46]. In order to overcome the challenge that a large training set is not available in the field of mammograms, we trained our CNN on LSVRC [49] which is a dataset containing more than 1 million labeled natural images first. As the natural images in the database are all with 3 channels, which are not in accordance with gray level medical images and increased the computing consuming in both training and testing stages. So they were transformed into gray scale ones with a simple projection method. And the training strategy mainly followed the strategy proposed by [33]. It is a supervised learning process forming hierarchical feature detectors, the learning rate of the training stage with LSVRC data was initialized at 0.01 and was divided by 10 when the validation error rate stopped improving with the current learning rate. The whole training process continued for 100 cycles through the natural image dataset. Then learning rate of training stage on breast mass images was initialized at 0.00001, and it changed as the strategy as

Table 2
Main parameters of each layer.

Name	Number of maps	ReLU	Pooling
Input	1	×	×
Conv1	96	√	√
Conv2	256	√	√
Conv3	384	√	×
Conv4	384	√	×
Conv5	256	√	√
Fc6	2048	√	×
Fc7	2048	√	×
Fc8	2	√	×

The **name**, **size**, **number of maps** and whether there were **ReLU** and **Pooling** units in each layer were illustrated in the table. In the **name** column, (Input) stands for the mass images while (Conv) and (Fc) represent convolution and fully-connected layer respectively. Values in the **size** and **number of maps** column are the size and number of output map in each layer. (√) and (×) in the last two columns illustrate whether there were **ReLU** or **Pooling** units in each layer.

that of the first training stage. And the second stage was executed 100 cycles. Meanwhile, main and details of the network structure are shown in Tables 1 and 2. Then we applied fine-tuning operation which takes the already learned model trained on LSVRC, adapted the architecture, and resumed training from the already learned model weights on our dataset of mass images. Specifically, breast dataset in this paper is made of 600 mass images of 227×227 which are extracted from DDSM [54]. According to the successful application of [60] on medical image and followed the diagnosis habits of the radiologist who we consulted from, instances we applied had been rotated angles of 90°, 180°, 270° to enlarge the dataset 3 times. Network parameters obtained from training on LSVRC were set as initial values in the specific CNN for mass images and optimized the network according to mass dataset. The main parameters of each convolution architecture in this network are given in Table 2. Size of feature maps in each layer are determined by both sliding window size and striding of sliding window in previous layers. And the number of feature maps or the length of features in each layer is resulted from number of different types of convolution and pooling kernels.

Stochastic gradient descent (SGD) [50] was employed to optimize our network which is very simple and efficient in the training process. The training point at each iteration was selected at random. Then the derivative of the loss term for that training sample was computed resulting in a gradient vector. And parameters were incrementally updated by moving toward the local minima in the direction of the gradient. The most important operation is computing derivative of the objective function, which is obtained by an application of the chain rule known as back-propagation. Generally speaking, a CNN model contains several of all these blocks above, forming directed acyclic graph (DAG). The DAG could be simplified as Fig. 3, where each output of corresponding block $(\boldsymbol{f}_1, \boldsymbol{f}_2, \boldsymbol{f}_3, \cdots\cdots \boldsymbol{f}_l)$ is described as $\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3, \cdots\cdots \boldsymbol{x}_l$, and the parameters of each layer was $\boldsymbol{w}_1, \boldsymbol{w}_2, \boldsymbol{w}_3, \cdots\cdots \boldsymbol{w}_l$. So the derivative of $\boldsymbol{w}_l$ in the loss layer is expressed in the form of back-propagation chain rule as follow.

$$\frac{dz}{d(vec\ \boldsymbol{w}_l)^T} = \frac{dz}{d(vec\ \boldsymbol{x}_L)^T} \frac{d\ vec\ \boldsymbol{x}_L}{d(vec\ \boldsymbol{x}_{L-1})^T} \cdots \frac{d\ vec\ \boldsymbol{x}_{l+1}}{d(vec\ \boldsymbol{x}_l)^T} \frac{d\ vec\ \boldsymbol{x}_l}{d(vec\ \boldsymbol{w}_l)^T} \tag{4}$$

To evaluate this CNN model, the loss of network was calculated by the standard cross entropy between the predicted probability distribution over two types of labels for each image and the ground truth distribution which equaled to the logarithmic loss. We used the backpropagation algorithm to calculate the gradient with respect to the parameters of the model and trained the network with stochastic gradient descent (SGD) by minimizing the loss as described above. To obtain a better result of this fine-tuning

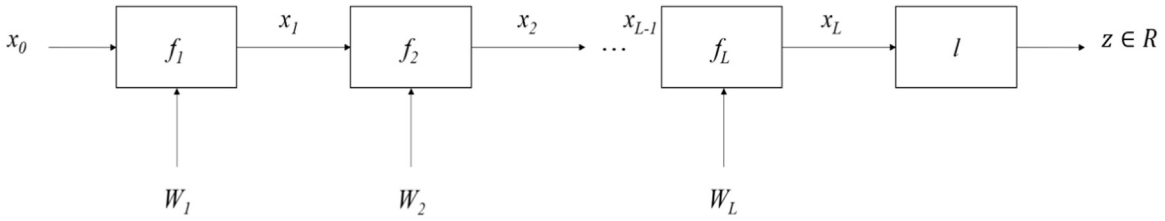


Fig. 3. The directed acyclic graph (DAG) representation of a CNN. Different layers in the structure were described by operation of weights and input successively.

operation and to avoid the oscillatory of loss function, we applied a strategy that the learning rate was reduced gradually during the whole process. Specifically, the learning rate is set to a smaller value if reduction of loss function is less than the threshold after several times of iterations. In addition, the initial value of learning rate is set up as a small one.

After obtaining the fine-tuned CNN, the hierarchical features could be extracted through the feed-forward model. For example, an instance is input into this structure to be handled with various methods of different layers. As a result of each operation, output of each layer is presented in a modality of feature maps which contained a lot of channels, and these are actually the hierarchical features we applied in the paper. Meanwhile, in both training and testing processes we used some basic functions of the toolbox named matconvnet [61].

2.2. Decision mechanism

The features we chose for training the classifiers which divide the input images into two types were extracted from two different layers of this network. Concretely speaking, they were from the layers of Conv5 and Fc7, both of which were in the form of column vector.

Then two linear SVM classifiers based on the hierarchical features were trained on training dataset and tested on test dataset to predict a test image was whether benign or malignant. The SVM we preferred was the most basic one in LIBSVM [51] toolbox, resulting in fast calculations in both training and testing procedures. And all its parameters were chosen as the defaults without the time consuming procedure of parameter selection. Comparing with the results of kernel SVM experiments and *k*-fold cross validation operation, there was a trade-off between efficiency and effectiveness of the performance when we chose the basic linear SVM classifier. As were shown in Fig. 5 and Table 3, we applied the SVM with RBF kernel and spent plenty of time doing experiments on selecting the optimal value of C and Gamma which are widely believed as the most important parameters of RBF kernel SVM in LIBSVM. Concretely speaking, we set both C and Gamma to a substantial range of $2^{-15}, 2^{-14}, \cdots, 2^{0}, \cdots, 2^{14}, 2^{15}$ and summarized the results. As it was shown in Fig. 5, two coordinate axes represented the values of C and Gamma which were in the form of power-of-2 while the legend value stood for classification accuracy. The best performance was 97.0% ($C = 2^{5}$, Gamma $= 2^{-12}$) which was a little bit better than that (96.7%) of the linear SVM with default parameters. However, it took us 77,180 s to do parameters selection experiments and it was more than ten thousand times of the experiment with default values. In addition, we applied *k*-fold cross validation in our linear SVM experiments. From the results in Table 3, we could clearly see that the time consuming *k*-fold cross validation did not improve the performance obviously either. So we still selected the linear SVM setting all parameters as default values for its far much higher efficiency and a good enough performance.

We propose a strategy to combine outcomes of these two classifiers. As shown in Fig. 4, in the testing process, if the outcomes of two classifiers consist with each other, we take the outcomes as correct judgments and add them to a subset of final result which is named as *result1*. Otherwise, the test images causing inconsistent outcomes are joined to a dataset called uncertain set. To decide which type each instance in the testing set

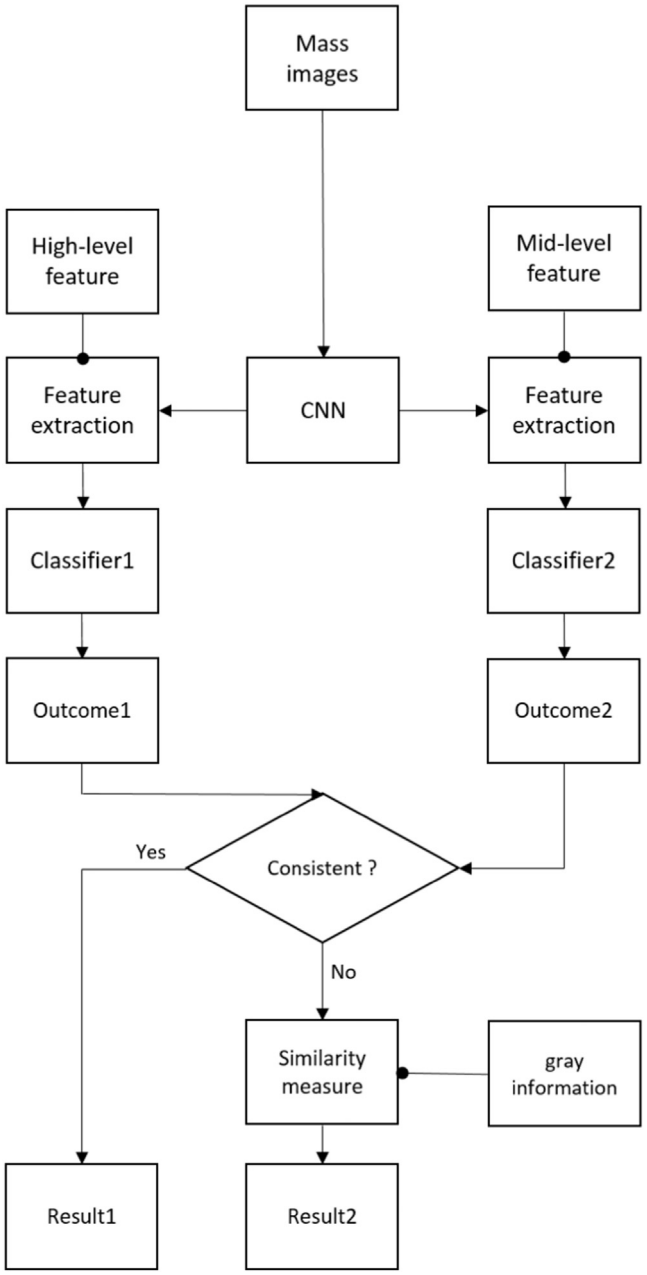


Fig. 4. The flow chart of the test process. As illustrated from the top down in this figure, high-level and middle-level features of a test image were extracted from the fine-tuned network before. Then these features were classified by two classifiers in a two-step decision mechanism.

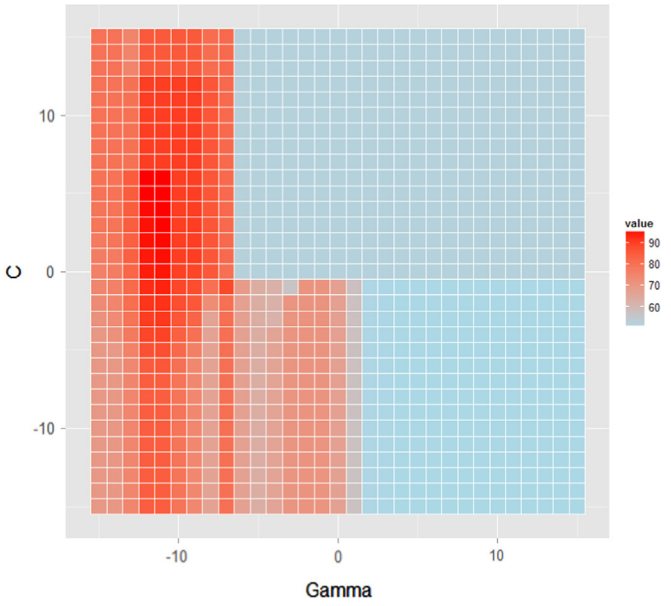


Fig. 5. Performance of kernel SVM with different C and Gamma.

Table 3
Classification accuracy of k-fold cross validation.

k (k-fold)	Classification accuracy (%)
2	96.7
3	97.1
4	97.3
5	97.6
6	97.5
7	97.6
8	97.5

belongs to, we use the original gray information to calculate closeness of these instances to benign and malignant ones in the training set. In the process of closeness calculation, benign and malignant images in the training dataset were clustered into several subclasses respectively, with the help of which we could obtain the cluster centers and number of each subclass in both two types for the next similarity measure.

In procedure of clustering, instances in training set are clustered by hierarchical K-means method [52]. In this method, images could be separated into several subclasses with unbalanced scales. Owing to the hierarchical method, the native data structure of dataset is kept by these clustering centers and unbalanced scales, both of which play important roles in the whole mechanism. According to the result of closeness measure, the uncertain set is divided into two types, forming *result2*, the other subset of final result. Consequently, the final result contains *result1* and *result2* is obtained.

The similarity to each kind of mass images is defined as follows. Here, instance in the uncertain set is $x_{U_i}, i=1,2,\ldots m$; clusters of benign and malignant ones in training set is $c_{B_j}, j=1,2,\ldots n$ and $c_{M_j}, j=1,2,\ldots n$ respectively; numbers of instances in subclass are $N_{B_j}, j=1,2,\ldots n$ in benign data and $N_{M_j}, j=1,2,\ldots n$ in malignant data; $|x_{U_i}-c_{B_j}|$ and $|x_{U_i}-c_{M_j}|$ were both euclidean distance of uncertain instance x_{U_i} to center of clustering. The similarity of one instance in the uncertain set to benign ones is:

$$S_{U_{iB}} = \frac{1}{\sum_{j=1}^{n} \frac{N_{B_j}}{N_B} |x_{U_i} - c_{B_j}|}, i=1,2,\ldots m \tag{5}$$

And similarity of one instance in the uncertain set to malignant ones is:

$$S_{U_{iM}} = \frac{1}{\sum_{j=1}^{n} \frac{N_{M_j}}{N_M} |x_{U_i} - c_{M_j}|}, i=1,2,\ldots m \tag{6}$$

After having obtained the similarity value of each instance, we make the final decision for this subset obey the rules given below:

If $S_{U_{iB}} > S_{U_{iM}}$, the instance is considered as benign;
If $S_{U_{iB}} < S_{U_{iM}}$, the instance is considered as malignant;
If $S_{U_{iB}} = S_{U_{iM}}$ which is rare in our experiments. We consider the instance is benign in accordance with the statistics of American government that the risk of benign breast masses was almost 3 times bigger than that of malignant ones [53].

The deep structure framework is proposed for extracting global, local and detail symptoms of mass images, which helped to form a unified description of mass images. The decision mechanism imitated physicians' diagnosis process which mainly contained symptom comparisons and antidiastole. Deep features extracted by our CNN played important roles in the scheme as they were thought to be in accordance with cognitive principle of human brain, as they were better simulations of the impressions left by masses on doctors, respectively.

3. Datasets and metrics

In this section, datasets and evaluation methods for both proposed method and reference methods were introduced.

3.1. Datasets

The dataset we chose to train our CNN model is a subset of *Imagenet* which is named LSVRC, and it contains more than 1 million natural images. Being the most popular natural image databases in the field of deep learning, LSVRC has become the choice of many researchers. It provides enough instances with determined labels, and this is quite important in training a supervised deep model.

Besides, the dataset we chose to do fine-tuning on for our CNN and conduct experiments on was a subset from the Digital Database for Screening Mammography (DDSM) [54]. The database is provided by the University of South Florida. DDSM contains more than 2600 cases, and each case includes four images above breast, along with associated patient information (age at time of study, ACR breast density rating, subtlety rating for abnormalities, ACR keyword description of abnormalities) and image information (scanner, spatial resolution, etc.). In this database, images containing suspicious areas have associated pixel-level “ground truth” information about the locations and types of suspicious regions.

The mammograms in DDSM have been detected and labeled to generate a dataset of 600 images before the classification, of which 50% are benign and 50% are malignant. Meanwhile, all the mass images we chose were representative and challenging ones. The mass images were divided into both training and testing sets of same size, of each of these sets were 150 benign and 150 malignant ones. And each instance in the dataset is a gray scale image with the size of 227×227. The method with which we detected the mass region was proposed in our previous work [55]. In order to improve the performance of CNN, both the training and testing images had been normalized and whitened before they were input to the network. And these operations were widely applied in a few deep learning models. Instances in the dataset were subtracted by their mean and they were normalized to the range of [0, 1]

Table 4
Classification performance of different methods in literature.

Reference	Features	Database/size of dataset	Classification accuracy (%)
Rangayyan et al.	Compactness, Fourier descriptors, moment-based shape factors, and chord-length statistics	MIAS & Calgary/54	88.9
Panchal et al.	Gray level based features, BIRADS features, patient age and subtlety features with Auto-associator MLP	DDSM/200	91.0
Mavroforakis et al.	Textural features and fractal dimension analysis	DDSM/130	83.9
Rojas et al.	Spiculation, gradient orientation, mutual information, fuzziness of mass margins	DDSM & MIAS /319	81.0
Cheng et al. [59]	BoW and histogram similarity	Galactograms/23	80.7
Verma et al.	Density, mass shape, margin, abnormality assessment rank, patient age, subtlety value	DDSM/200	88.8
Ramirez-Villegas et al.	Wavelet packet energy, Tsallis entropy and statistical parameterization	MiniMIAS/400	93.8
Xie et al.	Gray level features and textural features	DDSM/300 MIAS/60	95.7 96.0
Beura et al.	Combination of DWT, GLCM and BPNN	MIAS/320	97.4
Wang et al.	Latent spatial and statistical marginal characteristics	DDSM/600	92.7
Wang et al.	Data with agglomerative hierarchical clustering	DDSM/464	91.4
Ours	Deep features from different layers	DDSM/600	96.7

according to [56]. Then we whitened normalized data with the method named PCA whitening by dividing the standard deviation of its elements.

3.2. Metrics

To evaluate the performance of our proposed framework, we used both objective and subjective evaluation. With the help of consideration from these two aspects, we could appraise our method more effectively.

The objective measures we chose were receiver operating characteristic (ROC) curves and classification accuracy with deviations, both of which were preferred in most papers for evaluation of classification methods.

The definition of ROC curve is:

$$sensitivity = TP/(TP+FN)$$

$$specificity = TN/(FP+TN)$$

Here *TP* stands for the true positive cases in detection results, and *TN* denotes the true negative cases. In addition, *FP* contains the false positive cases, and *FN* equals the false negative cases. In the figure of ROC curve, the ordinate and abscissa were sensitivity and specificity respectively. A larger area under this curve stands for a better classification performance.

Another objective evaluation was the accuracy with deviations. In the task of cataloging mass images, classification accuracy is of vital importance. Performance with high accuracy could provide the doctor with a lot of help in the diagnostic process, which was a matter of survival and cure rates for patients. Given all these above, the accuracy with deviations was applied in this paper both for evaluation of our framework and for comparison with traditional methods. The calculation of these evaluation metrics were given as follows:

$$accuracy = \frac{1}{N}\sum_{i=1}^{N}\frac{TP_i+TN_i}{TP_i+FN_i+FP_i+TN_i}$$

$$std = \frac{accuracy - mean(accuracy)}{N}$$

Here, *N* is the number of testing times. *TP*, *TN*, *FP* and *FN* are in the same senses as they are mentioned above.

Two subjective evaluations we chose were described next.

In order to show the performance of feature extraction more intuitively, we chose t-distributed stochastic neighbor embedding (t-SNE) [48] as one of the subjective evaluation methods. The t-SNE method visualizes high-dimensional deep features and original images by giving each data point or instance a location in a two-dimensional map. This method was proposed by Hinton in 2006, and it has been proved to be quite effective in evaluation of various kinds of features. The t-SNE map can be explained as that the more linearly separable points in the two-dimensional map, the better this feature will perform.

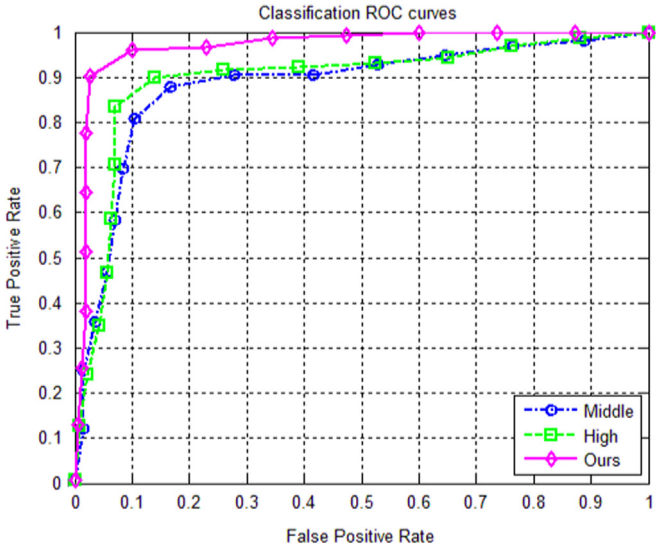


Fig. 6. ROC curves of middle-level feature, high-level feature based classification and our proposed method.

The last subjective metric was inspired by deconvolution network. To show what features of different levels are like, we applied deconvolution, activation and other operations on features in middle and high levels [57,58]. With the help of all these means, the rebuilding maps which were in the same size of input images assisted to show the emphasis of features on different levels. And we benefit from it to know whether these features could represent mass images in different scales.

4. Results and analysis

We compared the performance of the proposed framework with that of representative methods in the literature. Even the scales of datasets differed in these methods, they were in the same format. So the corresponding previously reported classification performance was still an effective standard to evaluate the improvement and setback of various methods. As shown in Table 4, a number of features were employed in these methods. It was obvious that frameworks with relatively good results used to apply 3 or more different kinds of features in the classification

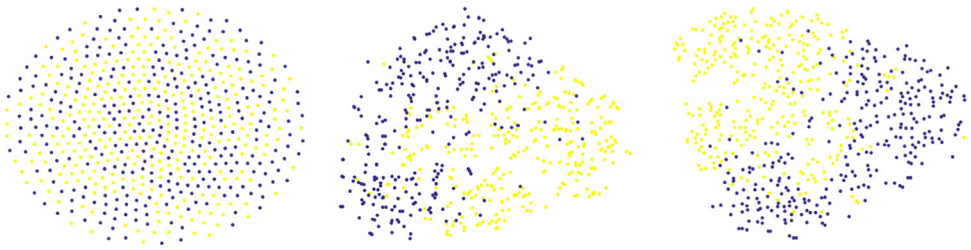


Fig. 7. The t-SNE maps of hierarchical features. From left to right are respectively original images, middle-level feature and high-level feature.

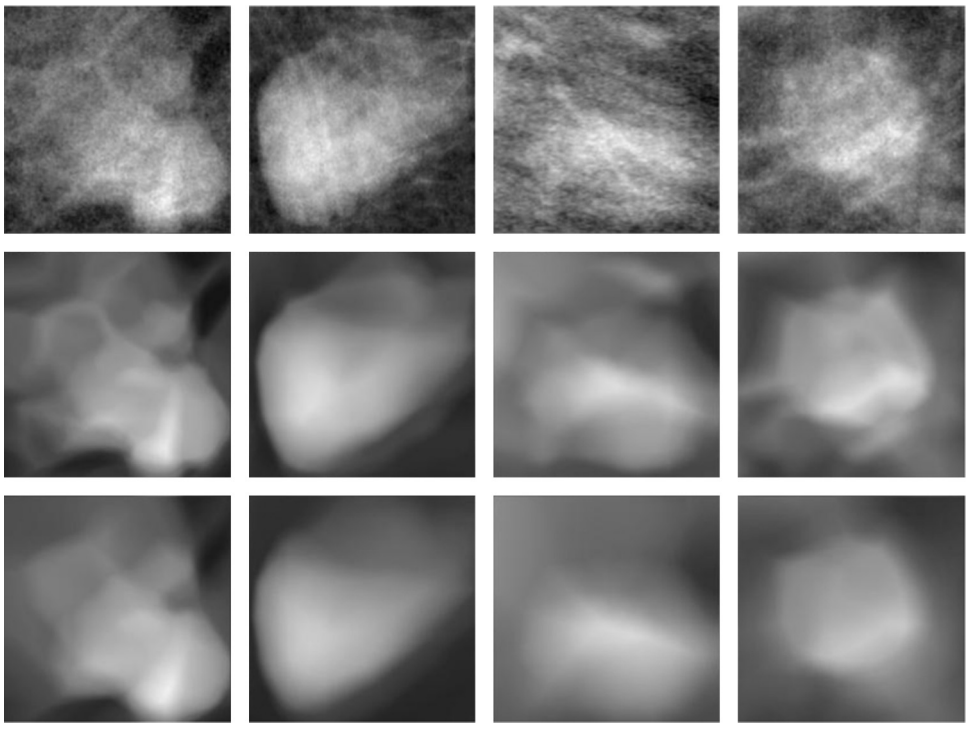


Fig. 8. Hierarchical features visualization of some benign instances. Each column represents an instance and from top to bottom are original images, middle-level feature and high-level feature.

task. As a result, the feature extraction process was complex in structure. Meanwhile, the challenge of making effective use of various types of features was inevitable in these frameworks, causing most of these researches made many efforts to find a fusion strategy for these features. However, methods with less complexity in feature extraction step did not seem to gain satisfactory results on a larger dataset which contained more than 200 instances. And results show clearly that a single kind of traditional feature was not sufficient for this task. The proposed framework took full advantage of deep features from a single CNN model, forming a unified feature extraction structure. From the comparison in the table, the proposed decision mechanism was not simple enough but effective to obtain good performance on a testing dataset of 300 instances. Both the scale of dataset applied and classification accuracy of our method were competitive among all these frameworks. Because it was a better simulation of physician's diagnosis procedure, our proposed method made more sense in terms of CAD.

As shown in Table 4 and Fig. 6, if the middle-level and high-level features were used individually for the classification task, performances were not that good. Once our decision mechanism joined, the whole framework could outperform a lot at both ROC curve and classification accuracy. Fig. 7 also demonstrates the ability of CNN model to extract discriminative features in a intuitive way. The method of t-SNE was used to form maps visualizing instances of the dataset in different levels. The evaluation criterion was that the more instances in the map were separable the better this feature performed. Here, the maps shown from left to right are the t-SNE maps of original images, middle-level and high-level features extracted from our CNN respectively. Different colors represented instances of different types. From all these maps, we could obviously notice that deep features improve the differentiation of two types of instances, which equaled that instances were much more separable after they were represented by deep features. It was of great help for classifiers to distinguish different kinds of instances.

To show what the features from different layers focused on, we adopted a deconvolution based method to visualize images hierarchically. As shown in Figs. 8 and 9, the first rows in both figures stood for original images of selected instances while the middle and bottom ones were middle-level and high-level features. As all the images came from the associated layers, they represented what levels emphasized on. And it was in evidence that maps from high-level feature paid more attention to the overall feel of mass images while that from middle-level feature and original image captured more local and details of the images. Meanwhile, the hierarchical expression was a progressive one during its

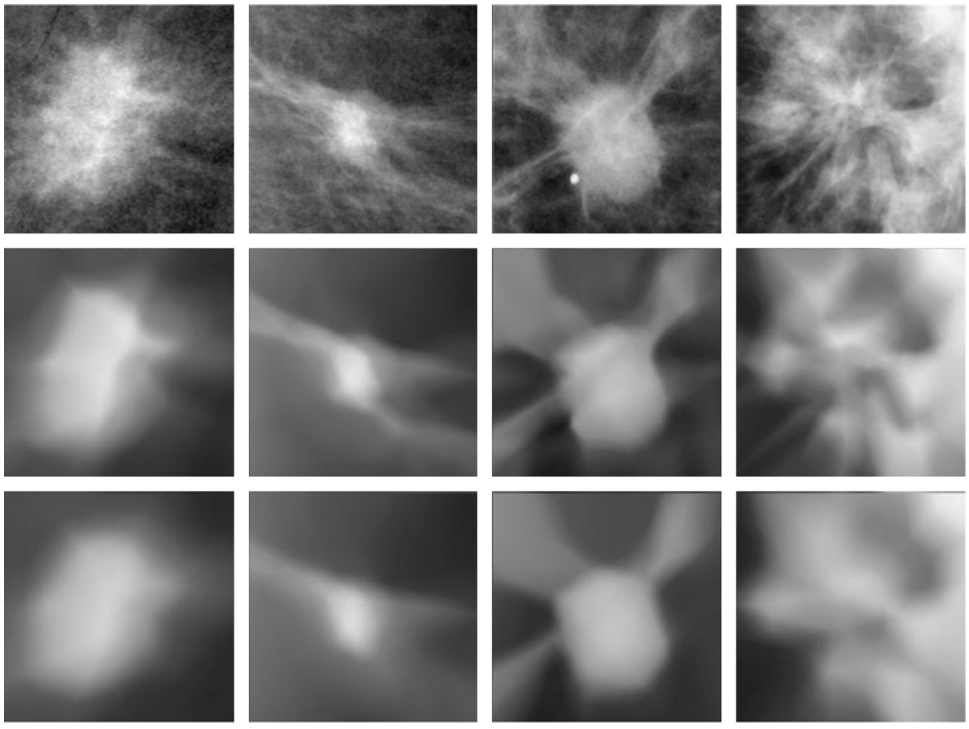


Fig. 9. Hierarchical features visualization of some malignant instances. Each column represents an instance and from top to bottom are original images, middle-level feature and high-level feature.

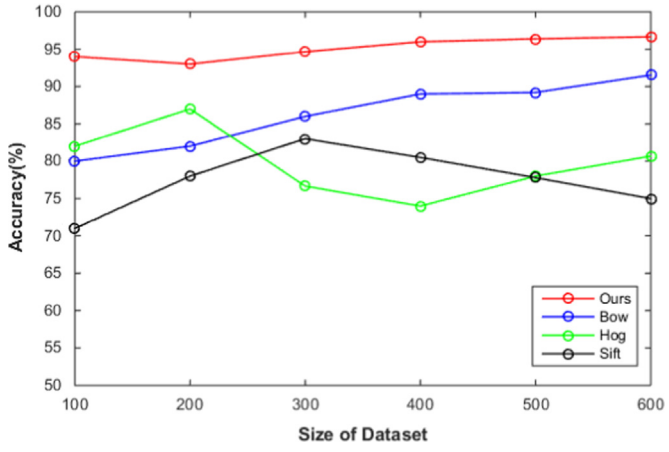


Fig. 10. Comparison of our scheme with other feature based methods on datasets of different sizes.

evaluation, which makes the feature representation more consistent with thinking activities in doctors' brain during the diagnosis.

In order to show the stability of our proposed algorithm when it comes to a situation that only a tiny database of breast mass images could be obtained, we execute our scheme on different scales of datasets and compare its outcomes with the state of art algorithm (BoW). In Fig. 10, points in the red curve stand for the classification accuracy of our method when it comes to datasets of various scales (100, 200, …600). And the blue curve is the performance of BoW-based scheme which is the state of art one in mass classification. The other two curves are also typical schemes which are histogram of oriented Gradient (HOG) and scale invariant feature transform (SIFT) based methods. From the two curves of different colors, we could significantly draw a conclusion that our proposed method outperforms the state of art one on both

Table 5
Comparison of our network with traditional ones.

Name	Size (MB)	Parameters	Time per image (ms)	Classification accuracy (%)
Caffe-ref	233	6.1e+07	2.97	92.0
VGG	528	1.4e+08	13.53	97.0
Ours	204	5.8e+07	1.10	96.7

stability and the accuracy of classification. Meanwhile it also demonstrates our method to be effective when large datasets are not available in the medical field which is one of the main challenges presented by this special issue.

In addition, in order to achieve the state of art results that were described in our paper, we performed various experiments to obtain the most suitable network structure for this task. We compared our network structure with the most widely used variations of Alex net at aspects of model size, number of parameters, time consuming and classification accuracy. These two models have been proved to be effective in a variety of tasks. From the contents of Table 5, we could see that our network achieved quite a competitive result with less storage space than Caffe-ref net. Though the classification accuracy of our net is a little bit worse than that of the VGG net, the time consuming, storage space and number of parameters are far much better.

5. Conclusion

In this paper, we proposed a novel mammographic masses classification framework. With the help of our previous work, the mass images were obtained from DDSM to form a dataset first. Instances in the dataset were whitened for preprocessing. Based on the dataset, we applied fine-tuning operation on the trained deep CNN model to acquire the feature extraction network for the

next procedures. All blocks in the deep structure were the most effective ones in the state-of-the-art papers. Then middle-level and high-level features were extracted from different layers of this network for training two linear SVM classifiers. Owing to the validity of deep features, the simplest linear classifiers are powerful enough to distinguish instances. Combining outcomes of these two classifiers, we got *result1* and the uncertain set. Meanwhile, the weighted similarity was calculated to decide labels of the uncertain set. Simulating the hierarchical structure of human brain and its data processing mechanism, the deep CNN model and features extracted from it was of great assistance in simulating multiscale impressions left by mass images on physician's brain. And the decision mechanism we proposed was a better imitation of doctors' symptom comparisons and antidiastole procedures in the diagnosis. Evaluated by two objective measures and compared with traditional effective methods, the proposed framework achieved better performance in classification accuracy. According to the results of two visualization strategies, the effectiveness of deep features in classification task for mass images has been verified intuitively. In addition, the experimental results demonstrated that the CNN model with adjustment for certain data was effective even the scale of specific database was not that large. In the future, we will make effort to find a better variation of CNN to help obtain more describable features and design a decision mechanism which is more like the real diagnosis procedure.

References

[1] A. Jemal, R. Siegel, E. Ward, et al., Cancer statistics, 2008, CA Cancer J. Clin. 58 (2) (2008) 71–96.

[2] K. Doi, Computer-aided diagnosis in medical imaging: historical review, current status and future potential, Comput. Med. Imaging Graph. 31 (4) (2007) 198–211.

[3] B. Van Ginneken, B.M. ter Haar Romeny, M. Viergever, Computer-aided diagnosis in chest radiography: a survey, IEEE Trans. Med. Imaging 20 (12) (2001) 1228–1241.

[4] Y. Jiang, R.M. Nishikawa, R.A. Schmidt, et al., Improving breast cancer diagnosis with computer-aided diagnosis, Acad. Radiol. 6 (1) (1999) 22–33.

[5] H.-P. Chan, K. Doi, C.J. Vybrony, et al., Improvement in radiologists' detection of clustered microcalcifications on mammograms: the potential of computer-aided diagnosis, Invest. Radiol. 25 (10) (1990) 1102–1110.

[6] R.M. Rangayyan, N.M. El-Faramawy, J.L. Desautels, et al., Measures of acutance and shape for classification of breast tumors, IEEE Trans. Med. Imaging 16 (6) (1997) 799–810.

[7] M.E. Mavroforakis, H.V. Georgiou, N. Dimitropoulos, et al., Mammographic masses characterization based on localized texture and dataset fractal analysis using linear, neural and support vector machine classifiers, Artif. Intell. Med. 37 (2) (2006) 145–162.

[8] S. Timp, C. Varela, N. Karssemeijer, Temporal change analysis for characterization of mass lesions in mammography, IEEE Trans. Med. Imaging 26 (7) (2007) 945–953.

[9] A. Rojas-Domínguez, A.K. Nandi, Development of tolerant features for characterization of masses in mammograms, Comput. Biol. Med. 39 (8) (2009) 678–688.

[10] S. Yoon, S. Kim, Mutual information-based SVM-RFE for diagnostic classification of digitized mammograms, Pattern. Recognit. Lett. 30 (16) (2009) 1489–1495.

[11] J.F. Ramirez-Villegas, D.F. Ramirez-Moreno, Wavelet packet energy, Tsallis entropy and statistical parameterization for support vector-based and neural-based classification of mammographic regions, Neurocomputing 77 (1) (2012) 82–100.

[12] D. Wang, L. Shi, P.A. Heng, Automatic detection of breast cancers in mammograms using structured support vector machines, Neurocomputing 72 (2009) 3296–3302.

[13] B. Verma, P. McLeod, A. Klevansky, A novel soft cluster neural network for the classification of suspicious areas in digital mammograms, Pattern. Recognit. 42 (9) (2009) 1845–1852.

[14] B. Verma, P. McLeod, A. Klevansky, Classification of benign and malignant patterns in digital mammograms for the diagnosis of breast cancer, Expert Syst. Appl. 37 (4) (2010) 3344–3351.

[15] Y. Wang, J. Li, X. Gao, Latent feature mining of spatial and marginal characteristics for mammographic mass classification, Neurocomputing 144 (2014) 107–118.

[16] S. Beura, B. Majhi, R. Dash, Mammogram classification using two dimensional discrete wavelet transform and gray-level co-occurrence matrix for detection of breast cancer, Neurocomputing 154 (2015) 1–14.

[17] W. Xie, Y. Li, Y. Ma, Breast mass classification in digital mammography based on extreme learning machine, Neurocomputing (2015).

[18] N. Cristianini, J. Shawe-Taylor, H. Lodhi, Latent semantic kernels, J. Intell. Inf. Syst. 18 (2–3) (2002) 127–152.

[19] Gabriella Csurka, L. Fan, et al., Visual categorization with bags of keypoints, In: Proceedings of the Workshop on Statistical Learning in Computer Vision, ECCV, vol. 1, no. 1–22, 2004.

[20] Robert Fergus, et al., Learning object categories from Google's image search, in: Proceedings of the Tenth IEEE International Conference on Computer Vision, ICCV, Vol. 2, 2005, pp.1816–1823.

[21] U. Avni, H. Greenspan, E. Konen, et al., X-ray categorization and retrieval on the organ and pathology level, using patch-based visual words, IEEE Trans. Med. Imaging 30 (3) (2011) 733–746.

[22] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, Nature 521 (7553) (2015) 436–444.

[23] Y. Bengio, A.C. Courville, Deep learning of representations, Handbook on Neural Information Processing, vol. 49, 2013.

[24] G.E. Hinton, R.R. Salakhutdinov, Reducing the dimensionality of data with neural networks, Science 313 (5786) (2006) 504–507.

[25] Yann LeCun, Koray Kavukcuoglu, Clément Farabet, Convolutional networks and applications in vision, in: Proceedings of IEEE International Symposium on Circuits and Systems (ISCAS), 2010, pp. 253–256.

[26] A. Ng, Sparse autoencoder, CS294A Lecture Notes, vol. 72, 2011.

[27] Ruslan Salakhutdinov, Geoffrey E. Hinton, Deep boltzmann machines, In: Proceedings of the 12th International Conference on Artificial Intelligence and Statistics, 2009, pp. 448–455.

[28] Ilya Sutskever, Geoffrey E. Hinton, Graham W. Taylor, The recurrent temporal restricted boltzmann machine, Adv. Neural Inf. Process. Syst. (2009) 1601–1608.

[29] K. Fukushima, Neocognitron: a self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position, Biol. Cybern. 36 (4) (1980) 193–202.

[30] Y. LeCun, L. Bottou, Y. Bengio, et al., Gradient-based learning applied to document recognition, Proc. IEEE 86 (11) (1998) 2278–2324.

[31] Le Cun, B. Boser, et al., Handwritten digit recognition with a back-propagation network, Adv. Neural Inf. Process. Syst. (1990).

[32] Y. Bengio, P. Lamblin, D. Popovici, et al., Greedy layer-wise training of deep networks, Adv. neural Inf. Process. Syst. 19 (2007) 153.

[33] Alex Krizhevsky, Ilya Sutskever, Geoffrey E. Hinton, Imagenet classification with deep convolutional neural networks, Adv. Neural Inf. Process. Syst. (2012) 1097–1105.

[34] C. Farabet, C. Couprie, L. Najman, et al., Learning hierarchical features for scene labeling, IEEE Trans. Pattern. Anal. Mach. Intell. 35 (8) (2013) 1915–1929.

[35] Ross Girshick, et al., Rich feature hierarchies for accurate object detection and semantic segmentation, In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 580–587.

[36] Viren Jain, et al., Supervised learning of image restoration with convolutional networks, In: Proceedings of the 11th IEEE International Conference on Computer Vision, ICCV, 2007, pp. 1–8.

[37] Viren Jain, Sebastian Seung, Natural image denoising with convolutional networks, Adv. Neural Inf. Process. Syst. (2009) 769–776.

[38] M. Helmstaedter, K.L. Briggman, S.C. Turaga, et al., Connectomic reconstruction of the inner plexiform layer in the mouse retina, Nature 500 (7461) (2013) 168–174.

[39] Dan Ciresan, et al., Deep neural networks segment neuronal membranes in electron microscopy images, Adv. Neural Inf. Process. Syst. (2012) 2843–2851.

[40] D.C. Cireşan, A. Giusti, L.M. Gambardella, et al., Mitosis detection in breast cancer histology images with deep neural networks, MICCAI (2013) 411–418.

[41] W. Zhang, R. Li, H. Deng, et al., Deep convolutional neural networks for multi-modality isointense infant brain image segmentation, Neuroimage 108 (2015) 214–224.

[42] C.F. Cadieu, H. Hong, D.L. Yamins, et al., Deep neural networks rival the representation of primate IT cortex for core visual object recognition, PLoS Comput. Biol. 10 (12) (2014) e1003963.

[43] V. Mnih, K. Kavukcuoglu, D. Silver, et al., Human-level control through deep reinforcement learning, Nature 518 (7540) (2015) 529–533.

[44] Xavier Glorot, Antoine Bordes, Yoshua Bengio, Deep sparse rectifier neural networks, in: Proceedings of the International Conference on Artificial Intelligence and Statistics, 2011, pp. 315–323.

[45] George E. Dahl, Tara N. Sainath, Geoffrey E. Hinton, Improving deep neural networks for LVCSR using rectified linear units and dropout, In: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2013, pp. 8609–8613.

[46] J. Donahue, Y. Jia, O. Vinyals, et al., Decaf: a deep convolutional activation feature for generic visual recognition, arXiv preprint arXiv:1310.1531, Oct. 2013.

[47] D.H. Hubel, T.N. Wiesel, Receptive fields, binocular interaction and functional architecture in the cat's visual cortex, J. Physiol. 160 (1) (1962) 106.

[48] L. Van der Maaten, G. Hinton, Visualizing data using t-SNE, J. Mach. Learn. Res. 9 (2579–2605) (2008) 85.

[49] Jia Deng, et al. Imagenet: a large-scale hierarchical image database, In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2009, pp. 248–255.

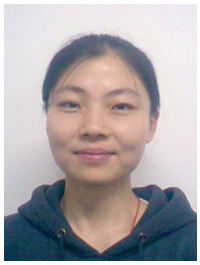
[50] Léon Bottou, Large-scale machine learning with stochastic gradient descent, In: Proceedings of COMPSTAT'2010, Physica-Verlag HD, 2010, pp. 177–186.

[51] C.-C. Chang, C.-J. Lin, LIBSVM: a library for support vector machines, ACM Trans. Intell. Syst. Tech. 2 (3) (2011) 27.

[52] Andrea Vedaldi, Brian Fulkerson, VLFeat: an open and portable library of computer vision algorithms, In: Proceedings of the 18th ACM International Conference on Multimedia, 2010, pp. 1469–1472.
[53] N. Howlader, A. Noone, M. Krapcho, et al., SEER cancer statistics review, 1975–2008, National Cancer Institute, Bethesda, MD, 2011.
[54] Digital Database for Screening Mammography (DDSM), University of South Florida, 2004.
[55] Y. Wang, D. Tao, X. Gao, et al., Mammographic mass segmentation: embedding multiple features in vector-valued level set in ambiguous regions, Pattern Recognit. 44 (9) (2011) 1903–1915.
[56] Adam Coates, Andrew Y. Ng, Honglak Lee, An analysis of single-layer networks in unsupervised feature learning, In: Proceedings of International Conference on Artificial Intelligence and Statistics, 2011, pp. 215–223.
[57] M.D. Zeiler, R. Fergus, Visualizing and understanding convolutional networks, Comput. Vision–ECCV (2014) 818–833.
[58] Mathew D. Zeiler, et al., Deconvolutional networks, In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2010, pp. 2528–2535.
[59] Erkang Cheng, et al., Mammographic image classification using histogram intersection, In: Proceedings of the 2010 IEEE International Symposium on Biomedical Imaging: From Nano to Macro, 2010, pp. 197–200.
[60] H.R. Roth, L. Lu, J. Liu, Improving Computer-aided detection using convolutional neural networks and random view aggregation, IEEE Trans. Med. Imaging (2015), http://dx.doi.org/10.1109/TMI.2015.2482920.
[61] A. Vedaldi, K. Lenc, MatConvNet-convolutional neural networks for MATLAB, arXiv preprint arXiv:1412.4564, 2014.

Zhicheng Jiao was born in 1990. He is a Ph.D. candidate at Xidian University. His research interests include intelligent information processing.

Xinbo Gao (M'02-SM'07) received the B.Eng., M.Sc. and Ph.D. degrees in signal and information processing from Xidian University, China, in 1994, 1997 and 1999 respectively. From 1997 to 1998, he was a research fellow in the Department of Computer Science at Shizuoka University, Japan. From 2000 to 2001, he was a postdoctoral research fellow in the Department of Information Engineering at the Chinese University of Hong Kong. Since 2001, he joined the School of Electronic Engineering at Xidian University. Currently, he is a Professor of Pattern Recognition and Intelligent System, and Director of the VIPS Lab, Xidian University. His research interests are computational intelligence, machine learning, computer vision, pattern recognition and wireless communications. In these areas, he has published 5 books and around 150 technical articles in refereed journals and proceedings including IEEE Transactions on Image Processing, the IEEE Transactions on Circuits and Systems for Video Technology, the IEEE Transactions on Neural Networks, and the IEEE Transactions on Systems, Man, and Cybernetics. He is on the editorial boards of several journals including Signal Processing (Elsevier), and Neurocomputing (Elsevier). He served as general chair/co-chair or program committee chair/co-chair or PC member for around 30 major international conferences. Now, he is a Fellow of IET and Senior Member of IEEE.



Ying Wang received the B.Sc., M.Sc. and doctor degrees in signal and information processing from the Xidian University, Xi'an, China, in 2003, 2006, and 2010 respectively. She is now an associate professor of Signal and Information Processing in Xidian University. Her research interests include medical image analysis, pattern recognition and computer-aided diagnosis.

Jie Li received the B.Sc., M.Sc. and Ph.D. degrees in Circuit and System from Xidian University, China, in 1995, 1998 and 2005 respectively. Since 1998, she joined the School of Electronic Engineering at Xidian University. Currently, she is a Professor of Xidian University. Her research interests include computational intelligence, machine learning, and image processing. In these areas, she has published over 30 technical articles in refereed journals and proceedings including IEEE TCSVT, IJFS etc.