Journal of the Institute of Industrial Applications Engineers Vol.8, No.4, pp.117–121, (2020.10.25)
DOI: 10.12792/JIIAE.8.117 Online edition: ISSN 2187-8811 Print edition: ISSN 2188-1758

Paper

Diagnosis of Breast Cancer from Mammogram Images Based on CNN

LIN DONG*† Member, KOHEI INOUE† Member

(Received October 1, 2020, revised December 1, 2020)

Abstract: Breast cancer has become the most common malignant tumor with the highest incidence of death in women. The MIBCAD (Medical Image Based Computer-Aided Diagnosis) system currently in use has a low diagnostic accuracy rate of only 85%. Furthermore, this system has major limitations for image processing of mammogram. To address these issues, this paper proposed a breast cancer diagnosis method based on an improved CNN (Convolutional Neural Networks). To avoid the image overfitting problem, transfer learning and data augmentation methods were used. The image classification accuracy was improved by using different CNN structures and changing the classifier type. Our results showed that the classification accuracy of the model reached 91.4%, which was significantly improved compared with the existing MIBCAD system.

Keywords: Mammography, Convolutional Neural Networks, Transfer Learning, Image Classification, Data Augmentation

1. Introduction

According to the Japan National Cancer Center, breast cancer is the most common cancer among women and the fifth most common cause of death from cancer. Therefore, the early diagnosis of breast tumors has become an important issue. Early diagnosis of breast cancer has higher requirements for doctors' professional standards. Mammography is the most common method of performing early screening for breast cancer. This method is inexpensive and causes less pain to the patient and clearly shows the breast tissue structure. Doctors use the images of the breast to determine if a lesion is present. However, the accuracy of this method of diagnosis depends on the doctor's prior experience, and due to the differences in the level of diagnosis and prior experience between doctors, misdiagnosis and omission can easily occur. Another major reason for misdiagnosis is the fatigue of the doctor who has to read the mammography for a long time, which affects his or her own judgment. Recent studies have shown that MIBCAD (Medical Image Based Computer-Aided Diagnosis) system is widely used to detect and diagnose breast cancer to improve doctors' efficiency.

Recent studies have shown that MIBCAD (Medical Image Based Computer-Aided Diagnosis) system is widely used to detect and diagnose breast cancer to improve doctors' efficiency. Studies have shown that the use of MIBCAD can help inexperienced doctors diagnose breast lesions with sensitivity from 62% to 85% and help experienced physicians diagnose breast lesions with sensitivity from 77% to 85% [1]. This shows that MIBCAD can indeed assist doctors in diagnosing lesions in the field of image diagnosis.

* Corresponding: dolly8060@hotmail.com
† Department of Communication Design Science, Kyushu University
4-9-1, Shiobaru, Minami-ku, Fukuoka 815-8540, Japan

In a previous study, it was found that methods of image classification of mammogram based on computer-aided diagnosis were mainly based on the traditional classification methods of artificial feature extraction (e.g., Zhang et al. [2]). In the past few years, with the development of deep learning, CNN (convolutional neural networks) have achieved excellent achievements in computer vision such as face recognition and handwritten character recognition, (e.g., Wang et al. [3] and Maitra et al. [4]). Recently, CNN have also been gradually used for medical image classification. Araújo et al. [5] used a CNN to classify H&E-stained breast cancer pathology images into benign and malignant tumors on the mammography dataset provided by the Israel Institute of Technology, achieving an accuracy rate of 88.3%. In an earlier study (Bayramoglu et al. [6]), it was verified that they achieved an accuracy rate of 83% in classifying mammogram images using magnification-independent deep learning based on the dataset published by Spanhol et al [7]. Taken together, these studies suggest that their mammogram image classification models have a good performance on the non-public datasets. Furthermore, the accuracy of their model is not satisfactory.

In the present study, we investigated how to improve the performance of CNN models on the publicly available dataset. By introducing transfer learning on the basis of deep learning and by adjusting the network structure and parameters on the basis of AlexNet, in the present study we achieved high accuracy on the target dataset.

The rest of this paper is organized as follows. Section 2 proposes a breast cancer diagnosis method based on CNN. Section 3 shows experimental results, where the proposed method is compared with the state-of-the-art method for breast cancer diagnosis. Section 4 discusses the results and related future studies. Finally, Section 5 concludes this paper.

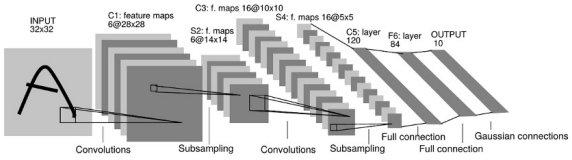


Figure 1: Architecture of LeNet-5 [8]

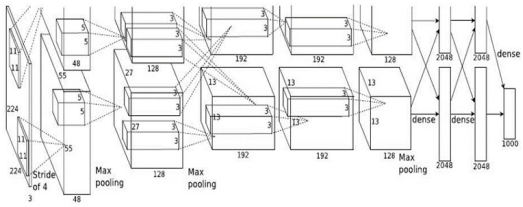


Figure 2: Architecture of AlexNet [9]

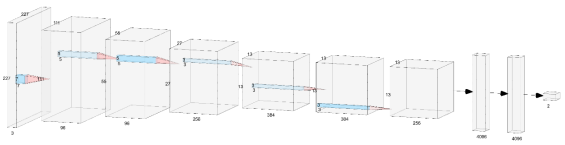


Figure 3: Architecture of the proposed model

2. Method

In this section, we propose a CNN model for breast cancer diagnosis.

2.1 Convolutional Neural Networks Recently, CNN have become a hot topic in the research field because of the high accuracy achieved in the field of image recognition, which is why we chose CNN.

In 1998, Y. LeCun [8] and others published a paper that established the modern structure of CNN, and later CNNs were perfected on its basis. They designed a multi-layered artificial neural network, named LeNet-5, which can classify handwritten numbers. Figure 1 shows the architecture of LeNet-5, which includes convolutional layers, pooling layers, and fully connected layers.

2.2 AlexNet network model Because of the good performance of AlexNet in Figure 2 by Krizhevsky et al. [9], in the present study we used the AlexNet neural network model to do image classification of mammogram. A typical CNN consists of an input layer, convolutional layers, pooling layers, fully connected layers, and an output layer. The AlexNet network model used in this paper consists of six convolutional layers and two fully connected layers as shown in Figure 3. The activation functions ReLU (Rectified Linear Unit) and LRN (Local Response Normalization) are included in each convolutional layer.

2.2.1 Input layer The input layer is responsible for loading the images and generating an output vector as input to the convolutional layer. The input to this model is a $227 \times 227 \times 3$ block, and all three channels of image information are used as inputs.

2.2.2 Convolutional layer In CNN, the convolutional kernel is the core of the network. The convolutional kernels are translated on a two-dimensional plane, and each element of the kernel is multiplied by the corresponding position of the convolved image, and then summed. By constantly moving the convolution kernel, we have a new image that consists entirely of the result of summing the products of the convolution kernels at each position. An important feature of the convolution algorithm is that the original signal features can be enhanced and the noise reduced by the convolution algorithm. Different convolution kernels can be used to extract different features of the image. In this study, I varied the size of the convolutional kernels in the convolutional layer for the purpose of improving the network performance. In the first convolutional layer, 96 convolutional kernels of size 7×7 with a step size of 2 were used. In the second convolutional layer, 96 convolutional kernels of size 5×5, step size 2, and padding =1 were used. In the third convolutional layer, 256 convolutional kernels of size 5×5 with a step size of 1 were used. For the fourth, fifth and sixth convolutional layers, 384 kernels of 3×3 with a step of 1 were used. Although changing the size and ordering of the convolution kernels increased the running time of the program, the advantage is that it improved the accuracy of image classification.

2.2.3 Pooling layer The pooling layer is usually followed by the convolutional layer, and the feature map is downsampled according to certain downsampling rules. The function of downsampling mainly has two points: 1) Reduce the dimensionality of the feature map; 2) Maintain the scale-invariant characteristics of the feature to a certain extent, and improve the performance and robustness of the algorithm. There are two common downsampling rules: mean-pooling and max-pooling. In this paper, a Max-pooling approach was used. The pooling size of the second, third and sixth pooling layers were all 3×3, and the step size was 2.

2.2.4 Fully connected layer Each neuron in the fully connected layer is connected to all neurons in the preceding layer. The model has 2 fully connected layers. The first fully connected layer used 4096 neurons in the 256 (6×6) feature maps obtained through convolution and downsampling to perform a full connection, and the number of output feature map units was 4096. For the second fully connected layer, the number of input feature map units was 4096 and the number of output feature map units was 4096.

2.2.5 Output layer In the present study, the output layer used the SVM (Support Vector Machine) classifier function instead of Softmax classifier function to classify the input images.

SVM is a commonly used classifier for binary classification of data based on supervised learning, widely used in data analysis, pattern recognition, and data regression. Since the dataset in this study had limited samples, SVM was used to classify the data. The classification principle used is to find the best hyperplane to divide the dataset. Good classification results can be achieved with a limited sample by this way. In the process of finding the best hyperplane, slack variables and penalty factors are added to

improve the generalization of the model and to reduce the possibility of overfitting. The slack variables allow the classification hyperplane to misclassify a portion of the sample. The penalty factor, on the other hand, acts as a regularization to control the complexity of the model. The above methods are used to improve the generalization of the model.

SVM can be broadly classified into linearly separable SVMs and linearly inseparable SVMs. Among them, the linearly inseparable SVM treats a linearly inseparable problem by mapping a low-dimensional linearly inseparable problem to a high-dimensional space using a kernel function, thus turning it into a linearly separable problem. This study replaced the Softmax classifier function in the original output layer with an SVM classifier function. The specific operation was mainly to replace the cross entropy loss function used by the Softmax classifier function with a loss function in the form of hinge loss of the SVM classifier function.

The SVM classifier eventually finds the N-dimensional space of the segmentation hyperplane H as the following function:

$$H : g(x) = \omega^T x + b = 0, \tag{1}$$

In Eq.(1), the ω represents the weights and b represents the bias values. The partition hyperplane H that separates the positive and negative samples is the optimal hyperplane that the SVM looks for. The best hyperplane is the hyperplane that is the farthest away from both positive and negative samples and has a strong ability to generalize to unknown samples. In the binary classification problem, two hyperplanes will be assumed, which are the hyperplane H_1 passing through the support vector of the positive samples and parallel to the hyperplane H as shown in Eq.(2), and the hyperplane H_2 passing through the support vector of the negative samples and parallel to the hyperplane H as shown in Eq.(3).

$$H_1 : g(x) = \omega^T x + b = 1, \tag{2}$$

$$H_2 : g(x) = \omega^T x + b = -1. \tag{3}$$

According to Eq.(2) and Eq.(3), the distance between the support vector and the optimal hyperplane should be $\frac{1}{|\omega|}$. So to find the maximum distance is to find the minimum value of $|\omega|$ and to find

$$\min\left(\frac{1}{2}\omega^T\omega\right). \tag{4}$$

At the same time, in order to prevent overfitting, a small part of the samples are allowed to be classified incorrectly. It is not only allowed that the predicted labels of some points are inconsistent with the true labels, but also the amount of inconsistent data is minimized. It can be expressed as

$$\min\left(c\sum_{n=1}^{N}\varepsilon_n\right). \tag{5}$$

Among them, N represents the number of samples that are classified incorrectly, and ε_n represents the distance between the incorrectly classified point on the hyperplane and the correct classification. The function of the parameter c is to limit the number of samples with incorrect classification. Adjusting the parameter c can determine whether the model pays more attention to the margin or the number of samples with incorrect classification, so that the accuracy of the model is kept within a certain range. When c is increased, the penalty for misclassification will increase, and ε_n will become smaller, which means that the classification is more stringent. When c is decreased, the penalty for misclassification will be reduced, and ε_n will become larger, which means greater fault tolerance. Combining Eq.(4) with Eq.(5), we get the Loss function of the SVM:

$$\min\left(\frac{1}{2}\omega^T\omega + c\sum_{n=1}^{N}\varepsilon_n\right) \;\; s.t. \;\; \omega^T x_n t_n \geq 1 - \varepsilon_n \;\; \forall n \tag{6}$$

where x_n is the sample being classified and t_n is the expected output corresponding to that sample. Hinge Loss function is a loss function used to train the classifier. For a two-class classifier, the formula of Hinge Loss function is

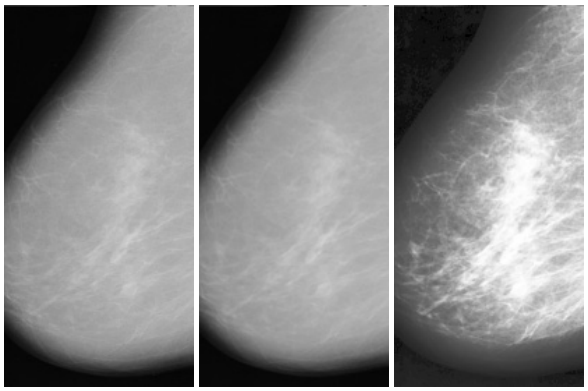
$$loss = \max\left(1 - \omega^T x_n t_n, 0\right). \tag{7}$$

For the fixed ω and b, the ε_n in Eq.(6) is also a fixed value, which is the result of the Hinge Loss function. Because the zero area of the Hinge loss function corresponds to the normal samples of non-support vectors, all normal samples do not participate in the determination of the optimal hyperplane. In this way, the dependence on the number of training samples is greatly reduced, and training efficiency is improved. Through Eq.(6) and Eq.(7), the Hinge Loss function form of the SVM was defined by Eq.(8).

$$\min\left(\frac{1}{2}\omega^T\omega + c\sum_{n=1}^{N}\max\left(1 - \omega^T x_n t_n, 0\right)\right). \tag{8}$$

Based on the three formulas above, the loss function of the SVM used in this paper can be viewed as the sum of the L2-normalization and hinge loss function.

2.3 Transfer learning The training of CNN's parameters requires large labeled datasets. Since medical images contain a lot of private information about individual patients, the disclosure of medical images requires the permission of each patient. Therefore, it is a challenge to collect large-scale datasets that are publicly available with the permission of the patient. On the other hand, due to the uneven level of doctors, high-quality labeled medical images are also scarce. Till now, for breast cancer diagnosis, large labeled datasets are lacking. Therefore, this paper adopted a transfer learning approach to address this issue. The basic idea of transfer learning is to pre-train a CNN on an existing large dataset and then transfer to the target dataset for training and fine-tuning using the weights of that pre-trained CNN as initialization weights. The reason why transfer learning is feasible is that the first few layers of



(a) (b) (c)

Figure 4: (a) cropped image, (b) denoised image, (c) enhanced image

CNN learn generic features of the datasets, such as points, lines, colors, and other underlying features, and the last few layers learn specific features of the datasets. In this paper, we pre-trained on an ImageNet dataset (consisting of more than 1.2 million natural images and 1000 different classes), and the resulting model's parameters were used as initialization parameters for the model, which were then transferred to the target dataset for training. Furthermore, we used a global fine-tuning strategy to optimize the model.

2.4 Dataset and Data augmentation In this paper, we just verified and tested the performance of the proposed model on existing mini mammogram dataset from MIAS (mammography image analysis society). The resolution of all mammograms in this dataset was 1024×1024, with 208 normal images and 114 abnormal images (63 benign tumors and 51 malignant tumors). Therefore, there were 322 mammograms in total. These mammograms were used directly by doctors so that our image classification program was also run directly on this dataset.

The left and right sides of each image in the mini mammogram dataset have black areas that do not affect image recognition, so their resolution was cropped to 1024×512. The image was denoised by median filtering, in order to improve the signal-to-noise ratio of the image. The image was then enhanced by histogram equalization. This was to highlight the desired features of the image. The results were shown in Figure 4. In order to enhance the robustness of the neural network and avoid overfitting, an adequate amount of data input was required. Therefore, the rotation method was used to expand the dataset. Each image was rotated by -90°, 90° and 180° around the origin, which made the existing dataset three times the original dataset.

2.5 Training method We randomly selected 70% of the dataset as a training set to train the neural network model, while the remaining 30% of the dataset was used as a test set to evaluate the performance of the neural network model. The number of epoch was 1000, and the batch size was the size of the training set.

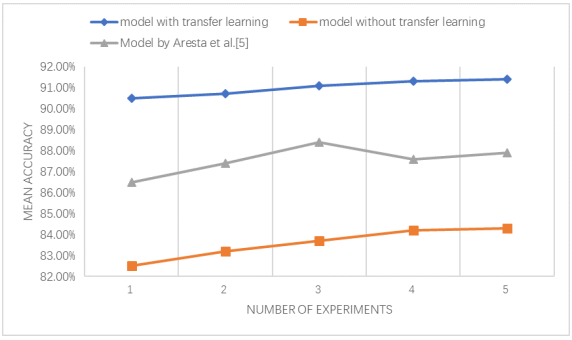


Figure 5: Comparison of the classification accuracy of different models

Table 1: Comparison of accuracy between different models

Training model	Final mean accuracy
Our model with transfer learning	**91.4**%
Model by Aresta et al. [5]	87.9%
Our model without transfer learning	84.3%

Table 2: Comparison between SVM classifier and Softmax classifier

Classifier	Mean accuracy	Training time in every batch
SVM	**91.4**%	**958**s
Softmax	88.9%	973s

3. Results

The Image classification program was run on a desktop computer (Intel CPU Core i7-8700; NVIDIA GeForce RTX 2080Ti; VENGEANCE LPX Series memory module). The CPU frequency was 3.2 GHz. The video memory was GDDR6 11 GB and the size of the memory module was 24 GB (one 8 GB and one 16 GB). The experimental model in this study was based on the Tensorflow framework.

First in this study, the classification model was compared before and after the introduction of transfer learning. At the same time, the performance of the breast cancer classification model of Araújo et al. [5] on the dataset in this study was used as a benchmark to compare and verify the performance of our classification model. The results of the experiments are shown in Figure 5 and Table 1. The mean accuracy of the experimental results was the mean of ten random assignment dataset experiments.

It is obvious from Figure 5 and Table 1 that after the introduction of transfer learning, the classification accuracy of our classification model was greatly improved and compared to that of Araújo et al. [5].

We then compared the training time and accuracy of the model using the SVM classifier and the Softmax classifier. The result of comparison is shown in Table 2.

This experiment shows that the training time using the SVM classifier is shorter than that using the Softmax classifier. This is because for the SVM classifier, when the probability of a class is greater than 0.9, it means that the classification is correct and the classifier no longer processes the

samples that are already classified correctly, thus drastically reducing the training time and increasing the generalization capacity of the network. For Softmax classifier, however, the loss function continues to compute until the probability of correct classification is close to 1, resulting in an increase in training time.

4. Discussion

The results showed that the mammography classification system based on the AlexNet network model proposed obtained an accuracy of 91.4% on the test set, which improved the accuracy by 3.10% compared to the study by Araújo et al. [5]. The best explanation for the high accuracy from our models was that we used AlexNet network to classify. Certainly, the accuracy of our model was also significantly improved by introducing transfer learning. Taken together, this paper established a mammography classification system based on the AlexNet network model on the basis of transfer learning. This effectively improved the classification performance of the system and provided experience for other medical small data image classification systems.

In future studies, we will work to improve the following issues. Since the dataset in the paper has not considered the effect of the imbalance of positive and negative samples on the convolutional neural network, this suggests that we should have verified our convolutional neural network model on more datasets. In addition, our model can only distinguish whether it is breast cancer or not. In the future, the classification system must be able to accurately classify specific types of breast cancer and not just be able to classify whether it is breast cancer or not.

5. Conclusion

In this study, we proposed a CNN-based classification model for mammogram image classification. We adopted the AlexNet network model and modified the size of the convolution kernel. At the same time, in the output layer, we replaced the Softmax classifier with the SVM classifier, because the SVM classifier greatly shortens the training time. At the same time, the SVM classifier can quickly converge to the optimal value, which improves the classification accuracy of the classification model. By introducing transfer learning and data augmentation, overfitting is effectively prevented and classification accuracy is significantly improved. Experiments show that the CNN classification model proposed in this paper achieves an average accuracy of 91.4% on the target data set, which achieves a high average classification accuracy.

References

[1] C. Balleyguier, K. Kinkel, J. Fermanian and M. Sebastien, "Computer-aided detection (CAD) in mammography: does it help the junior or the senior radiologist," *European Journal of Radiology*, Vol. 54, No. 1, pp.90-96, 2005. DOI: 10.1016/j.ejrad.2004.11.021

[2] Y. Zhang, B. Zhang and W. Lu, "Breast cancer histological image classification with multiple features and random subspace classifier ensemble," *Knowledge-Based Systems in Biomedicine and Computational Life Science*, pp. 20-58, 2013. DOI: 10.1007/978-3-642-33015-5_2

[3] W. Wang, J. Yang, J. Xiao, et al., "Face recognition based on deep learning," *International Conference on Human Centered Computing*, pp. 812-820, 2014. DOI: 10.1007/978-3-319-15554-8_73

[4] D. S. Maitra, U. Bhattacharya and S. K. Parui. "CNN based common approach to handwritten character recognition of multiple scripts," *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*, pp.1021-1025, 2015. DOI: 10.1109/ICDAR.2015.7333916

[5] T. Araújo, G. Aresta, E. Castro and R. José, "Classification of breast cancer histology images using convolutional neural networks," *PloS one*, Vol. 12, No. 6, p. e0177544, 2017. DOI: 10.1371/journal.pone.0177544

[6] N. Bayramoglu, J. Kannala and J. Heikkilä, "Deep learning for magnification independent breast cancer histopathology image classification," *2016 23rd International conference on pattern recognition (ICPR)*, pp. 2440-2445. DOI: 10.1109/ICPR.2016.7900002

[7] F. A. Spanhol, L. S. Oliveira, C. Petitjean and H. Laurent, "A dataset for breast cancer histopathological image classification," *IEEE Transactions on Biomedical Engineering*, Vol. 63, No. 7, pp. 1455-1462, 2015. DOI: 10.1109/TBME.2015.2496264

[8] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, Vol.86, No. 11, pp. 2278-2324, 1998. DOI: 10.1109/5.726791

[9] A. Krizhevsky, I. Sutskever and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *International Conference on Neural Information Processing Systems*, pp. 1097-1105, 2012. DOI: 10.1145/3065386

Lin Dong (Member) received B.E. degree from Harbin Institute of Technology of China in 2018. He is currently a graduate student in Kyushu University. His research interests include deep learning and image processing.

Kohei Inoue (Member) received B.Des., M.Des. and D.Eng. degrees from Kyushu Institute of Design in 1996, 1998 and 2000, respectively. He is currently an Associate Professor in Kyushu University. His research interests include pattern recognition and image processing.