Breast Cancer Detection Using Transfer Learning in Convolutional Neural Networks

Shuyue Guan
Department of Biomedical Engineering
The George Washington University Medical Center
Washington DC, USA
frankshuyueguan@gwu.edu

Murray Loew
Department of Biomedical Engineering
The George Washington University Medical Center
Washington DC, USA
loew@gwu.edu

***Abstract*—In the U.S., breast cancer is diagnosed in about 12% of women during their lifetime and it is the second leading reason for women's death. Since early diagnosis could improve treatment outcomes and longer survival times for breast cancer patients, it is significant to develop breast cancer detection techniques. The Convolutional Neural Network (CNN) can extract features from images automatically and then perform classification. To train the CNN from scratch, however, requires a large number of labeled images, which is infeasible for some kinds of medical image data such as mammographic tumor images. A promising solution is to apply transfer learning in CNN. In this paper, we firstly tested three training methods on the MIAS database: 1) trained a CNN from scratch, 2) applied the pre-trained VGG-16 model to extract features from input mammograms and used these features to train a Neural Network (NN)-classifier, 3) updated the weights in several final layers of the pre-trained VGG-16 model by back-propagation (fine-tuning) to detect abnormal regions. We found that method 2) is ideal for study because the classification accuracy of fine-tuning model was just 0.008 higher than that of feature extraction model but time cost of feature extraction model was only about 5% of that of the fine-tuning model. Then, we used method 2) to classify regions: benign vs. normal, malignant vs. normal and abnormal vs. normal from the DDSM database with 10-fold cross validation. The average validation accuracy converged at about 0.905 for abnormal vs. normal cases, and there was no obvious overfitting. This study shows that applying transfer learning in CNN can detect breast cancer from mammograms, and training a NN-classifier by feature extraction is a faster method in transfer learning.**

Keywords—breast mass classification; transfer learning; deep learning; convolutional neural networks; mammogram; computer aided diagnosis; fine-tuning

I. INTRODUCTION

Breast cancer will be diagnosed among about 1 in 8 (or 12%) U.S. women during their lifetime [1] and it is the second leading reason for women's death in the U.S. [2]. Early mammographic detection based on computer-aided detection (CAD) methods can improve treatment outcomes for breast cancer and longer survival times for the patients [3]. Traditional CAD tools rely on manually extracted features, but they have a variety of drawbacks; for example, hand-crafted features tend to be domain specific, and the process of feature design can be tedious, difficult, and non-generalizable [4]. An alternative method for feature extraction is to learn features from whole images directly through a Convolutional Neural Network (CNN) [5], [6]. The CNN has performed well in many image classification tasks [7]. For example, in 2012, the AlexNet (a classical CNN model) won the ImageNet Challenge [8], which contains color images with 1000 classes. It achieved an accuracy of 83.6% for the top-5 error. To train the CNN from scratch, however, requires a large number of labeled images [9]: AlexNet was trained by using about 1.2 million labeled images [8]. Such a requirement often is infeasible for some kinds of medical image data such as mammographic tumor images because they are difficult to obtain, true positives are scarce in the datasets, and expert labeling is expensive [10]. A promising solution is to reuse as the feature extractor a pre-trained CNN model that has been trained with very large image datasets from other fields, or re-train (fine-tune) such a model using a limited number of labeled medical images [11]. This approach is also called transfer learning, which has been successfully applied to various computer vision questions [12]–[14]. In fact, some results of transfer learning are counterintuitive: previous studies for the pulmonary embolism and melanocytic lesion detection [11], [15] show that the features (connection weights in the CNN) learned from natural images could be transferred to medical images, even if the target images greatly differ from the pre-trained source images.

Currently, CNN has been applied to medical image classification in three major ways: 1) training CNN from scratch [16]–[18]; 2) using pre-trained CNN model to extract features from medical images [19]–[21] and 3) fine-tuning pre-trained CNN model on medical images [22]–[24]. In this study, we compared the three main techniques to detect breast cancer using the Mammographic Image Analysis Society (MIAS) mammogram database [25].

Previous studies have applied various machine learning methods for breast cancer/tumor detection using mammograms [26]. The MIAS and Digital Database for Screening Mammography (DDSM) [27] are the most commonly used public mammogram databases, and 10-fold cross validation is widely used to test trained models. Some studies used the traditional automatic feature extraction (not manual extraction) techniques, such as Gabor filter, fractional Fourier transform and Gray Level Co-Occurrence Matrix (GLCM), to obtain

features and then applied SVM or other classifier to do classification [28]–[32]. Neural networks were also used as classifiers [33], [34]. And some studies applied CNN to generate features from mammographic images [35]–[38]. Some of these studies used pre-trained CNN as applications of transfer learning. Few previous studies, however, presented results obtained by using only CNN for both feature generation and classification for breast cancer detection in mammograms. In our study, we used only one CNN; its front convolutional layers are responsible for feature generation and the back fully-connected (FC) layers are the classifier. Thus, the input for our CNN is mammographic images and its output are the (predicted) labels.

In this paper, we used mammographic images from the two databases: MIAS and DDSM. Firstly, we tested three training methods on MIAS: 1) trained a CNN from scratch, 2) applied the pre-trained VGG-16 model [39] to extract features from input images and used these features to train a Neural Network (NN)-classifier, 3) updated the weights in several last layers of VGG-16 model by back-propagation (fine-tuning) to detect abnormal regions. By comparison, we found that the method 2) is ideal for study. Secondly, we used method 2) to classify regions: benign vs. normal, malignant vs. normal and abnormal vs. normal from DDSM. We applied 10-fold cross validation to evaluate classification results. The validation accuracy curves converged, and there was no obvious overfitting.

Compared with other studies in this field, this study used a different pre-trained model, a simpler classification architecture and classifier, and used many more images for training. The results are competitive with prior works: our average accuracy is about 0.905 for abnormal vs. normal classifications and the AUC = 0.96. Our best model could reach 0.950 accuracy for abnormal vs. normal case.

II. Materials

A. Mammograohy Databases

Mammography is the process of using low-energy X-rays to examine the human breast for diagnosis and screening. There are two main angles to get the X-ray images: the cranio-caudal (CC) view and the mediolateral-oblique (MLO) view (Fig. 1). The goal of mammography is the early detection of breast cancer [40], typically through detection of masses or abnormal regions from the formed X-ray images. Usually, such abnormal regions are spotted by doctors or expert radiologists.

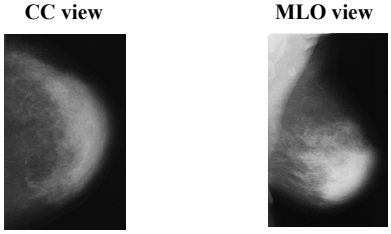

Fig. 1. Mammography in CC and MLO view

In this study, we used mammographic images from the two databases: Mammographic Image Analysis Society (MIAS) [25] and Digital Database for Screening Mammography (DDSM) [27]. The MIAS is an organization of UK research groups interested in the understanding of mammograms and has generated a database of digital mammograms. The MIAS database has 322 images including 102 abnormal and 220 normal samples. The locations and boundaries of these abnormal regions are given. The DDSM is another widely used resource by the U.S. mammographic image analysis research community. It is a collaborative effort between Massachusetts General Hospital, Sandia National Laboratories and the University of South Florida Computer Science and Engineering Department. The DDSM database contains approximately 2,620 cases in total: 695 normal cases, 1925 abnormal cases (914 malignant\cancers cases, 870 benign cases and 141 benign without callback) with locations and boundaries of abnormalities. Each case includes four images representing the left and right breasts in CC and MLO views.

B. Pre-trained model: VGG-16

For transfer learning, we applied the pre-trained VGG-16 model [39] in this study. The VGG-16 network was proposed by the Oxford Visual Geometry Group for the ImageNet Large-Scale Visual Recognition Challenge (ILSVRC) competition. This model is also known as one of typical deep convolutional networks. It was deeper and wider than the previous neural architectures. It mainly consists of five groups of convolution operations. Adjacent convolution groups are connected through max-pooling layers. Each group contains a series of 3×3-pixel convolutional layers. The VGG-16 model has 16 hidden layers in total, composed of 13 convolutional layers and 3 FC layers.

This pre-trained VGG-16 network was trained with about 1.3 million images (1000 classes) from ImageNet database [41] (ILSVRC-2012 competition), and it surpassed human-level performance on ImageNet [42], which achieved 7.5% top-5 error on ILSVRC-2012-Val and 7.4% top-5 error on ILSVRC-2012-Test in the competition.

III. Methods

In this study, we firstly downloaded mammographic images in MIAS and DDSM databases and cropped the Region of Interest (ROI) by given abnormal areas as ground truth information. We proposed three training methods including non-transfer learning and transfer learning in CNN.

A. Images Pre-processing

We downloaded all mammographic images in MIAS and DDSM databases from their official website. The images in MIAS are in PGM format, which can be read and processed by MATLAB directly. But images in DDSM are compressed in LJPEG format. To decompress and convert these images, we used the DDSM Utility [43]. We converted all images in DDSM to PNG format. MIAS describes abnormal regions by circular boundaries, and their center locations (X, Y) and radii values are contained in the documentation. DDSM describes the location and boundary of actual abnormality by chain-codes, which are recorded in OVERLAY files for each breast image containing abnormalities. The DDSM Utility also

provides the tool to read boundary information and display them for each image having abnormalities. Since the DDSM Utility tools run on MATLAB, we implemented all pre-processing tasks in MATLAB.

We used the ROIs instead of whole images to train neural networks. These ROIs are cropped rectangle-shape images and obtained by:

- For abnormal ROIs from images containing abnormalities, they are the minimum rectangle-shape areas surrounding the whole given ground truth boundaries.
- We firstly obtained abnormal ROIs. So, for normal ROIs, they are also rectangle-shape images and their size are about the average size of abnormal ROIs in the same database. For example, in DDSM, the average size of abnormal ROIs is 506.02×503.90 pixels, the decided cropping size for normal ROIs is 505×505 pixels. Their locations are randomly selected on normal breast areas. In this study, we cropped only one ROI from a whole normal breast image.

The sizes of abnormal ROIs vary with abnormality boundaries. Since the CNN requires all input images to be one specific size and usual inputs for CNN are RGB images (images in MIAS and DDSM are Grey images and the input of VGG-16 model requires RGB images), we resized the ROIs by resampling and made them to RGB (3-layer cubes) by duplication (Fig. 2).

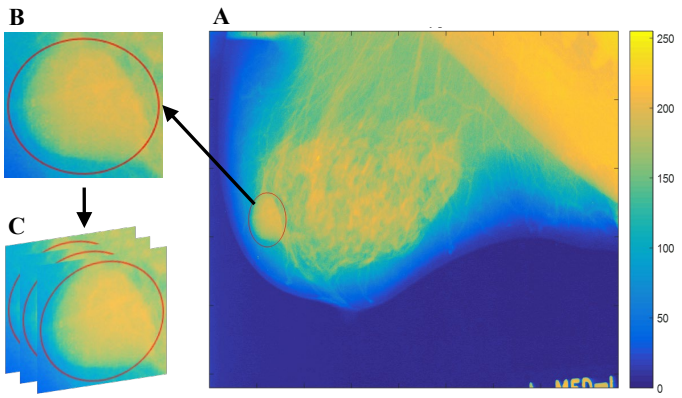

Fig. 2. (A) A mammographic image from MIAS in pseudo color (Parula [MATLAB]); (B) Cropped ROI by the given truth abnormality boundary; (C) Convert Grey to RGB image by duplication.

B. *To Train the Convolutional Neural Network (CNN) from Scratch (New-model)*

We built our own CNN in this part. The details about this CNN structure show in the TABLE I. It consists of three convolutional layers with max-pooling layers and one FC layer. The activation function for each layer is the ReLU function [44] except the last one for output, which is sigmoid function.

The notation Conv_3-32 means there are 32 convolutional neurons (units) and the filter size in each unit is 3×3-pixel (height×width) in this layer. MaxPool_2 means a max-pooling layer with size of filters is 2×2-pixel window, stride 2. And FC_64 means a fully-connected layer having 64 units. Dropout layer [45] randomly set a fraction rate of input units to 0 for the next layer at every updating during training; it could help the CNN avoid overfitting. The output layer uses a sigmoid function, which maps the output value to the range of [0, 1].

TABLE I. CNN ARCHITECTURE FOR TRAINING FROM SCRATCH

input: RGB image
Conv_3-32 + ReLU
MaxPool_2
Conv_3-32 + ReLU
MaxPool_2
Conv_3-64 + ReLU
MaxPool_2
FC_64 + ReLU (with Dropout = 0.5)
output (sigmoid): [0, 1]

C. *Transfer Learning: Features Extraction by Pre-trained VGG-16 network (Feature-model)*

The structure of CNN in transfer learning was combined the 13 convolutional layers in pre-trained VGG-16 model [39] with a simple FC layer (TABLE II).

TABLE II. CNN ARCHITECTURE FOR TRANSFER LEARNING

input: RGB image		
VGG-16	Conv block 1	Conv_3-64 + ReLU
		Conv_3-64 + ReLU
		MaxPool_2
	Conv block 2	Conv_3-128 + ReLU
		Conv_3-128 + ReLU
		MaxPool_2
	Conv block 3	Conv_3-256 + ReLU
		Conv_3-256 + ReLU
		Conv_3-256 + ReLU
		MaxPool_2
	Conv block 4	Conv_3-512 + ReLU
		Conv_3-512 + ReLU
		Conv_3-512 + ReLU
		MaxPool_2
	Conv block 5	Conv_3-512 + ReLU
		Conv_3-512 + ReLU
		Conv_3-512 + ReLU
		MaxPool_2
FC_256 + ReLU (with Dropout = 0.5)		
output (sigmoid): [0, 1]		

As shown in TABLE II, all the weights in 5 convolutional blocks (the blue background layers) were imported from the pre-trained VGG-16 model and not changed (or called weights frozen) during the training of this CNN. Only weights in the FC layer were randomly initialized and updated by training. Thus, such training process can be seen as that the VGG-16 extracts features from input image and then these features were used to train a FC NN-classifier.

D. Transfer Learning: Fine-tuning (Tuning-model)

The CNN structure for fine-tuning is the same structure as shown in TABLE II. One difference is in the training process – not all weights in the pre-trained model are fixed. During the fine-tuning training, the weights in the first 4 convolutional blocks (the darker blue background layers) were imported from the pre-trained VGG-16 model and frozen. The weights in the last convolutional blocks (Conv block 5), however, were updated by training. Another difference is that weights in the FC layer were imported from previous feature extraction training instead of random initialization. Weights in the last convolutional blocks were also imported from the pre-trained VGG-16 model. Therefore, no weight was randomly initialized in fine-tuning.

IV. RESULTS AND EVAULUATIONS

Our implementation of CNN was on the Keras API backend on TensorFlow [46]. The development environment for Python was Anaconda3.

A. Results for MIAS

We firstly tested the three CNN classification models – New-model, Feature-model and Tuning-model – on MIAS dataset. We randomly selected 95 ROI (cropped) images for each abnormal and normal case, and divided them into training and validation set by ratio 15:4. The label is binary, which "0" stands for normal and "1" for abnormal. Our training method (optimizer) was RMSprop [47] using default parameters provided in Keras, loss function was Binary Cross Entropy, updating metrics was Accuracy, batch size was 15 and the number of total epochs was set to be 500. For a CNN classifier, the input is the ROI image in size 421×421-pixel. Since the sigmoid function was used in the output layer, the predicted outcome from the CNN classifier is a value between 0 and 1. By default, the classification threshold is 0.5, meaning that if the value is less than 0.5 it will be considered as "0" (normal), otherwise it will be considered as "1" (abnormal).

1) New-model (non-transfer learning):

The result in Fig. 3 shows classification accuracy of the New-model for validation set. The blue curve is the accuracy after each epoch of training, and it was smoothed (the smoothing interval is about 20 epochs) to yield the red curve because we want to see its tendency as the number of epochs increased. One epoch means the model has been trained by all training data once. This result shows that the average accuracy is low (Max = 0.751) and the accuracy curve (blue) has not converged.

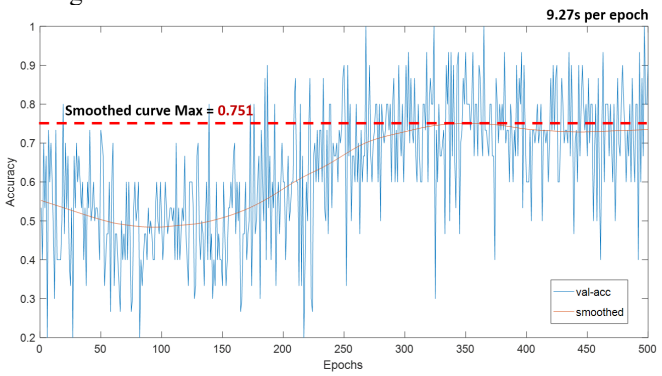

Fig. 3. Result of the New-model

2) Feature-model:

The result in Fig. 4 shows the average accuracy of the Feature-model converged at about 0.906 (also Max = 0.906) and the accuracy curve converged. The time cost for each epoch is about 14% of that of the New-model. Therefore, such comparison demonstrates that the performance of CNN in transfer learning is much better than training from scratch for breast cancer/tumor detection.

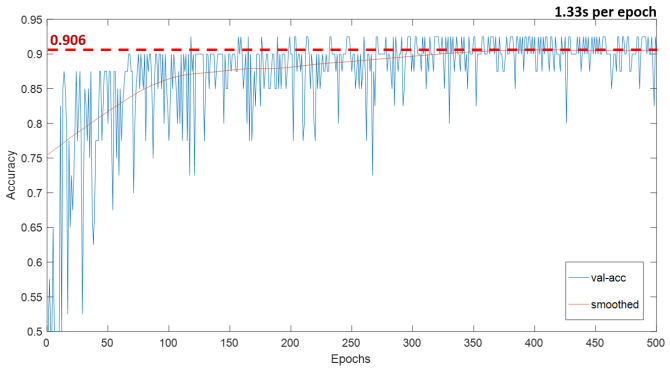

Fig. 4. Result of the Feature-model

3) Tuning-model:

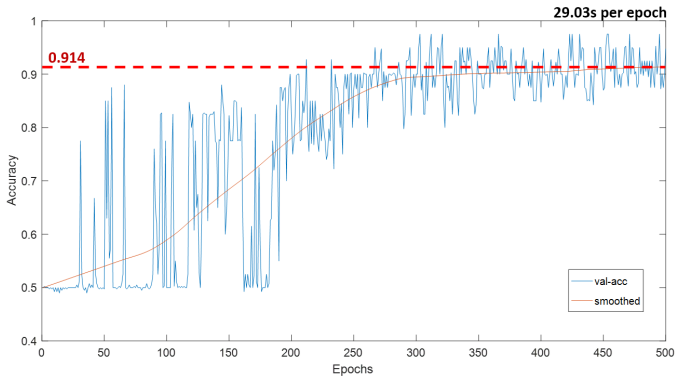

Fig. 5. Result of the Tuning-model

The result in Fig. 5 shows the average accuracy of the Tuning-model can reach a maximum of 0.914 and the accuracy curve also converged. Its performance is slightly improved

(about 0.88%) compared to the Feature-model. But the training time for each epoch is about 22 times that of training the classifier by only feature extraction.

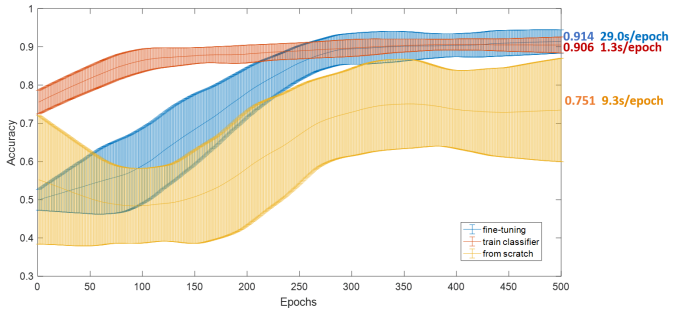

Fig. 6. Comparing of the three CNN classification models

Fig. 6 shows classification accuracy of the three models – New-model (yellow), Feature-model (red) and Tuning-model (blue) – on the MIAS validation set. The center line is smoothed accuracy (the smoothing interval is about 20 epochs) and width shows the departure of the mean. By comparison, training classifier by extracted features is the ideal method for study because its accuracy is very close to that of fine-tuning and the time cost is only about 5% of that of fine-tuning. But for real applications, fine-tuning is also feasible because we can have enough time (off-line) to train a very good model for implementation.

B. Results for DDSM

Second, we tested the performance of the Feature-model on the DDSM dataset. We used the same pre-processing as with MIAS to crop the DDSM images to create ROIs. Since the DDSM dataset has three main categories (normal, benign and malignant\cancer), we designed three classification experiments. The number of ROIs used for each experiment shows in TABLE III. All ROIs were randomly selected and shuffled in class sets. The label is binary, which "0" stands for Class 2 (normal) and "1" for Class 1(abnormal).

TABLE III. EXPERIMENTS FOR FEATURE-MODEL ON DDSM

Experiment	Class 1	# of ROIs	Class 2	# of ROIs
Exp.1	benign	800	normal	800
Exp.2	malignant	800	normal	800
Exp.3	abnormal[a]	1300	normal	1300

[a] The abnormal class contains 650 benign and 650 malignant ROIs.

The training method used the Feature-model, and evaluation was 10-fold cross validation. Our training method (optimizer) was Nadam [48] using default parameters (except the learning rate changed to 1e-4) provided in Keras, the loss function was Binary Cross Entropy, the updating metric was Accuracy, the batch size was 20 and the number of total epochs was set to be 500. For a CNN classifier, the input is the ROI image (of size 300×300-pixels) and the predicted outcome is a value between 0 and 1. By default, the discriminant threshold is 0.5. To compute the Receiver Operating Characteristic (ROC) curve, such threshold could be changed between 0 and 1.

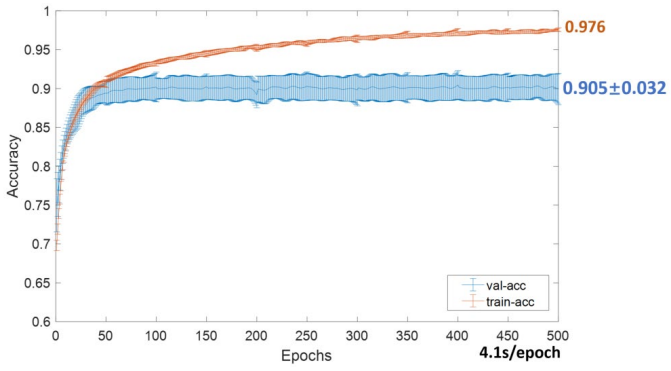

Fig. 7. Exp.3 result

Fig. 7 shows the result of Exp. 3. The red curve is training accuracy and blue is validation accuracy. The center line is average through 10-cross validation accuracy, and width shows the range of fluctuation in 10-cross validation. After 100 epochs, the validation accuracy converged at about 0.905, and there was no obvious overfitting. This classification accuracy matches the result on the MIAS database, which is about 0.906.

During the training in Exp. 3, the best (maximum) validation accuracy that the classification model reached was 0.950. Fig. 8 shows the ROC curve and AUC (area under curve) value for that model. Besides the average performance, the best situation is also important because during the CNN training, for each epoch, the validation accuracy may be changed but we could keep the best model for using.

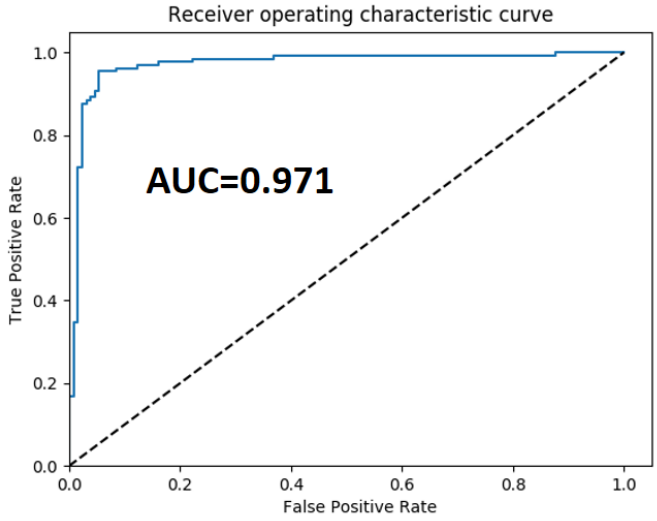

Fig. 8. The ROC curve of the best model (Acc=0.950) in Exp.3

TABLE IV. CLASSIFICATION RESULTS FOR FEATURE-MODEL ON DDSM

Experiment	Val_Acc	Tr_Acc	Best Val_Acc model	
			Val_Acc	*AUC*
Exp.1	0.909±0.044	0.983	0.975	0.993
Exp.2	0.912±0.035	0.988	0.956	0.980
Exp.3	0.905±0.032	0.976	0.950	0.971

Since Exp. 3 has more ROIs and its abnormal set includes half benign and half malignant ROIs, it is more representative than Exp. 1 and Exp. 2. TABLE IV shows results for the three experiments, where Val_Acc is the accuracy for validation set and Tr_Acc is accuracy for training set.

It is reasonable that Exp. 2 has the highest average validation accuracy of classification because a malignant mass could have more differences from normal tissue than a benign mass.

V. DISCUSSION

A. Comparsion of Related Studies

We reviewed several recent studies highly related to ours. These studies (TABLE V) applied transfer learning in CNN to detect breast cancer/abnormality based on mammogram.

By comparison with these studies, we used many more mammographic images for training and testing the CNN classifiers and a distinct pre-trained model. The main difference is about the classifier. Our one-FC layer NN-classifier has simpler architecture and could be integrated with pre-trained convolutional layers as one complete CNN. The average classification accuracy of our CNN (about 90.5%) for abnormal vs normal cases on mammograms and the AUC (about 0.96) are competitive to others'. And our best model could reach ACC = 95% for abnormal vs normal cases.

TABLE V. COMPARISON OF RELATED STUDIES

Main method	Validation (# of images)	Accuracy %	AUC
Pre-trained CNN on LSVRC datasets & Fine-tuning + Two-step decision[36]	2-fold cross (600)	(Ben-Mal) 96.7	–
Pre-trained CNN with hand crafted features + RF[37]	5-fold cross (410)	(Ben-Mal) 91 ± 0.02	0.76
Pre-trained AlexNet +Sparse MIL[35]	5-fold cross (410)	(Mal-nonMal) 90.00 ± 0.02	0.85
Pre-trained VGG-16 + one FC layer(Ours)	10-fold cross (2600)	(Abnorm-Norm) 90.5 ± 3.2	0.96

B. Problems and Future Works

There are two problems we found during this study: 1) the classification accuracy was much lower for benign (800) vs. malignant (800) ROIs, even for the best model (Max = 0.725). But we consider that the importance of benign vs. malignant is less than abnormal vs. normal because it is not the advantage or main purpose of mammography detection. 2) The classification accuracy by using mammogram of CC view is better than MLO view. We tested the Feature-model on benign (800) vs. normal (800) cases of CC view and MLO view. For the best classification accuracy, using CC view was 0.931 and using MLO view was 0.887. Their difference is small, but we did not find the reason or explanation yet.

For our future studies, we could try to recognize the abnormal areas in whole mammographic images. As the object detection with region proposal [49], by using the CNN, we could recognize the abnormalities on mammographic images and draw boundaries (or rectangle region proposals) on such areas automatically. These regions do not have to be 100% accuracy; they just provide another kind of reference for doctors to make decisions.

We could use other pre-trained models, and compare to their performances. In the research field of deep learning, VGG-16 appeared early but its depth (total number of layers is 23) is relatively shallow compared to new models, such as InceptionV3 (159 layers) [50], ResNet50 (168 layers) [51] and InceptionResNetV2 (572 layers) [52]. It will be interesting to see performances of breast cancer detection by using very deep CNNs.

VI. CONCLUSION

In this paper, we applied three CNN methods to detect breast cancer from mammograms. Training CNN from scratch is not feasible for limited number of labeled mammographic images. Using transfer learning in CNN is a promising solution for breast cancer detection. Our results show that the pre-trained CNN model (VGG-16) can automatically extract features from mammographic images, and a good NN-classifier can be trained by these features without providing hand-crafted features. Combining pre-trained CNN (VGG-16) with a one-FC NN-classifier can achieve average accuracy about 0.905 for classifying abnormal vs. normal cases in the DDSM database. In this study, the classification accuracy of the fine-tuning model is just 0.008 higher than that of the feature-extraction model but the time cost of the feature-extraction model is only about 5% of that of fine-tuning model. Therefore, this study shows that applying transfer learning in CNN can detect breast cancer from mammogram, and training a NN-classifier by feature extraction is a faster method in transfer learning.

REFERENCES

[1] C. E. DeSantis, S. A. Fedewa, A. Goding Sauer, J. L. Kramer, R. A. Smith, and A. Jemal, "Breast cancer statistics, 2015: Convergence of incidence rates between black and white women," *CA. Cancer J. Clin.*, vol. 66, no. 1, pp. 31–42, Feb. 2016.

[2] R. L. Siegel, K. D. Miller, and A. Jemal, "Cancer statistics, 2016," *CA. Cancer J. Clin.*, vol. 66, no. 1, pp. 7–30, Jan. 2016.

[3] V. M. Rao, D. C. Levin, L. Parker, B. Cavanaugh, A. J. Frangos, and J. H. Sunshine, "How Widely Is Computer-Aided Detection Used in Screening and Diagnostic Mammography?," *J. Am. Coll. Radiol.*, vol. 7, no. 10, pp. 802–805, Oct. 2010.

[4] D. Yi *et al.*, "Optimizing and Visualizing Deep Learning for Benign/Malignant Classification in Breast Tumors," *ArXiv170506362 Cs*, May 2017.

[5] S.-C. B. Lo, H.-P. Chan, J.-S. Lin, H. Li, M. T. Freedman, and S. K. Mun, "Artificial convolution neural network for medical image pattern recognition," *Neural Netw.*, vol. 8, no. 7–8, pp. 1201–1214, 1995.

[6] A. R. Jamieson, K. Drukker, and M. L. Giger, "Breast image feature learning with adaptive deconvolutional

networks," presented at the Medical Imaging 2012: Computer-Aided Diagnosis, 2012, vol. 8315, p. 831506.
[7] V. Sze, Y.-H. Chen, T.-J. Yang, and J. Emer, "Efficient Processing of Deep Neural Networks: A Tutorial and Survey," *ArXiv170309039 Cs*, Mar. 2017.
[8] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," in *Advances in Neural Information Processing Systems 25*, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2012, pp. 1097–1105.
[9] D. Erhan, P.-A. Manzagol, Y. Bengio, S. Bengio, and P. Vincent, "The Difficulty of Training Deep Architectures and the Effect of Unsupervised Pre-Training," 2009.
[10] H. C. Shin *et al.*, "Deep Convolutional Neural Networks for Computer-Aided Detection: CNN Architectures, Dataset Characteristics and Transfer Learning," *IEEE Trans. Med. Imaging*, vol. 35, no. 5, pp. 1285–1298, May 2016.
[11] N. Tajbakhsh *et al.*, "Convolutional Neural Networks for Medical Image Analysis: Full Training or Fine Tuning?," *IEEE Trans. Med. Imaging*, vol. 35, no. 5, pp. 1299–1312, May 2016.
[12] A. Sharif Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, "CNN features off-the-shelf: an astounding baseline for recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2014, pp. 806–813.
[13] H. Azizpour, A. Sharif Razavian, J. Sullivan, A. Maki, and S. Carlsson, "From generic to specific deep representations for visual recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2015, pp. 36–45.
[14] O. A. Penatti, K. Nogueira, and J. A. dos Santos, "Do deep features generalize from everyday objects to remote sensing and aerial scenes domains?," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2015, pp. 44–51.
[15] A. Esteva *et al.*, "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, vol. 542, no. 7639, pp. 115–118, Feb. 2017.
[16] J. M. Wolterink, T. Leiner, M. A. Viergever, and I. Išgum, "Automatic coronary calcium scoring in cardiac CT angiography using convolutional neural networks," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2015, pp. 589–596.
[17] Y. Pan *et al.*, "Brain tumor grading based on neural networks and convolutional neural networks," in *Engineering in Medicine and Biology Society (EMBC), 2015 37th Annual International Conference of the IEEE*, 2015, pp. 699–702.
[18] W. Shen, M. Zhou, F. Yang, C. Yang, and J. Tian, "Multi-scale convolutional neural networks for lung nodule classification," in *International Conference on Information Processing in Medical Imaging*, 2015, pp. 588–599.
[19] B. van Ginneken, A. A. A. Setio, C. Jacobs, and F. Ciompi, "Off-the-shelf convolutional neural network features for pulmonary nodule detection in computed tomography scans," in *2015 IEEE 12th International Symposium on Biomedical Imaging (ISBI)*, 2015, pp. 286–289.
[20] Y. Bar, I. Diamant, L. Wolf, S. Lieberman, E. Konen, and H. Greenspan, "Chest pathology detection using deep learning with non-medical training," in *Biomedical Imaging (ISBI), 2015 IEEE 12th International Symposium on*, 2015, pp. 294–297.
[21] F. Ciompi *et al.*, "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.*, vol. 26, no. 1, pp. 195–202, 2015.
[22] T. Schlegl, J. Ofner, and G. Langs, "Unsupervised pre-training across image domains improves lung tissue classification," in *International MICCAI Workshop on Medical Computer Vision*, 2014, pp. 82–93.
[23] G. Carneiro and J. C. Nascimento, "Combining multiple dynamic models and deep learning architectures for tracking the left ventricle endocardium in ultrasound data," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 11, pp. 2592–2607, 2013.
[24] R. Li *et al.*, "Deep learning based imaging data completion for improved brain disease diagnosis," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2014, pp. 305–312.
[25] J. Suckling *et al.*, "The Mammographic Image Analysis Society Digital Mammogram Database," *Exerpta Medica*, vol. 1069, pp. 375–378, 1994.
[26] K. Ganesan, U. R. Acharya, C. K. Chua, L. C. Min, K. T. Abraham, and K. H. Ng, "Computer-Aided Breast Cancer Detection Using Mammograms: A Review," *IEEE Rev. Biomed. Eng.*, vol. 6, pp. 77–98, 2013.
[27] M. Heath, K. Bowyer, D. Kopans, R. Moore, and W. P. Kegelmeyer, "The digital database for screening mammography," in *Proceedings of the 5th international workshop on digital mammography*, 2000, pp. 212–218.
[28] S. Khan, M. Hussain, H. Aboalsamh, and G. Bebis, "A comparison of different Gabor feature extraction approaches for mass classification in mammography," *Multimed. Tools Appl.*, vol. 76, no. 1, pp. 33–57, Jan. 2017.
[29] U. Raghavendra, U. Rajendra Acharya, H. Fujita, A. Gudigar, J. H. Tan, and S. Chokkadi, "Application of Gabor wavelet and Locality Sensitive Discriminant Analysis for automated identification of breast cancer using digitized mammogram images," *Appl. Soft Comput.*, vol. 46, pp. 151–161, Sep. 2016.
[30] S. Khan, M. Hussain, H. Aboalsamh, H. Mathkour, G. Bebis, and M. Zakariah, "Optimized Gabor features for mass classification in mammography," *Appl. Soft Comput.*, vol. 44, pp. 267–280, Jul. 2016.
[31] Y.-D. Zhang, S.-H. Wang, G. Liu, and J. Yang, "Computer-aided diagnosis of abnormal breasts in

mammogram images by weighted-type fractional Fourier transform," *Adv. Mech. Eng.*, vol. 8, no. 2, p. 1687814016634243, Feb. 2016.

[32] F. Narváez, J. Alvarez, J. D. Garcia-Arteaga, J. Tarquino, and E. Romero, "Characterizing Architectural Distortion in Mammograms by Linear Saliency," *J. Med. Syst.*, vol. 41, no. 2, p. 26, Feb. 2017.

[33] S. Wang, R. V. Rao, P. Chen, Y. Zhang, A. Liu, and L. Wei, "Abnormal Breast Detection in Mammogram Images by Feed-forward Neural Network Trained by Jaya Algorithm," *Fundam. Informaticae*, vol. 151, no. 1–4, pp. 191–211, Jan. 2017.

[34] R. Nithya and B. Santhi, "Classification of normal and abnormal patterns in digital mammograms for diagnosis of breast cancer," *Int. J. Comput. Appl.*, vol. 28, no. 6, pp. 21–25, 2011.

[35] W. Zhu, Q. Lou, Y. S. Vang, and X. Xie, "Deep Multi-instance Networks with Sparse Label Assignment for Whole Mammogram Classification," *ArXiv161205968 Cs*, Dec. 2016.

[36] Z. Jiao, X. Gao, Y. Wang, and J. Li, "A deep feature based framework for breast masses classification," *Neurocomputing*, vol. 197, pp. 221–231, Jul. 2016.

[37] N. Dhungel, G. Carneiro, and A. P. Bradley, "The Automated Learning of Deep Features for Breast Mass Classification from Mammograms," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016*, 2016, pp. 106–114.

[38] W. Borges Sampaio, E. Moraes Diniz, A. Corrêa Silva, A. Cardoso de Paiva, and M. Gattass, "Detection of masses in mammogram images using CNN, geostatistic functions and SVM," *Comput. Biol. Med.*, vol. 41, no. 8, pp. 653–664, Aug. 2011.

[39] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," *ArXiv14091556 Cs*, Sep. 2014.

[40] S. M. Friedewald *et al.*, "Breast Cancer Screening Using Tomosynthesis in Combination With Digital Mammography," *JAMA*, vol. 311, no. 24, pp. 2499–2507, Jun. 2014.

[41] O. Russakovsky *et al.*, "Imagenet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, 2015.

[42] K. He, X. Zhang, S. Ren, and J. Sun, "Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification," presented at the Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 1026–1034.

[43] A. Sharma, *DDSM Utility*. GitHub, 2015.

[44] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proceedings of the 27th international conference on machine learning (ICML-10)*, 2010, pp. 807–814.

[45] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting.," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, 2014.

[46] M. Abadi *et al.*, "TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems," *ArXiv160304467 Cs*, Mar. 2016.

[47] T. Tieleman and G. Hinton, "Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude," *COURSERA Neural Netw. Mach. Learn.*, vol. 4, no. 2, pp. 26–31, 2012.

[48] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," *ArXiv14126980 Cs*, Dec. 2014.

[49] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," *ArXiv150601497 Cs*, Jun. 2015.

[50] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the Inception Architecture for Computer Vision," *ArXiv151200567 Cs*, Dec. 2015.

[51] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," *ArXiv151203385 Cs*, Dec. 2015.

[52] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, "Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning," *ArXiv160207261 Cs*, Feb. 2016.