Breast Mass Lesion Classification in Mammograms by Transfer Learning

Fan Jiang
[1]Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China;
[2]Kunming University of Science and Technology, Kunming, China
fan.jiang@siat.ac.cn

Hui Liu
Kunming University of Science and Technology, Kunming, China
liuhui621@126.com

Shaode Yu and Yaoqin Xie
Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China
{sd.yu, yq.xie}@siat.ac.cn

ABSTRACT
Automatic classification of breast mass lesions in mammographic images remains an unsolved problem. This paper explored the technique of transfer learning to tackle this problem. It utilized the convolutional neural network (CNN) of GoogLeNet and AlexNet pre-trained on a large-scale visual database. The performance was evaluated a new dataset in terms of the area under the receiver operating characteristic curves (AUC). Results demonstrate that GoogLeNet (AUC=0.88) outperforms AlexNet (AUC=0.83) and other state-of-the-art traditional approaches in breast cancer diagnosis. The technique of transfer learning not only overcomes the unsatisfactory performance of traditional approaches, but also breaks the obstacle of limited samples for building deep CNNs.

CCS Concepts
• Computing methodologies~Biometrics

Keywords
Breast cancer; convolutional neural network; transfer learning; computer-aided diagnosis

1. INTRODUCTION
Breast cancer is the most common type of invasive cancer among women all over the world. In United States, estimated new cases of breast cancer is 0.25 million and estimated deaths is 0.04 million women in 2016 [1]. In developed countries, approximately one in eight women will develop this disease in the course of her life. A report reveals that breast cancer dramatically increases the health burden of developing countries [2].

Mammography is the routine imaging modality for breast cancer screening and functions as the currently gold standard for breast cancer programs. It produces high quality images to visualize the internal anatomy of the breast and quantify breast density, playing a vital important role in early detection and diagnosis of breast cancer and related disease and [3].

Double-reading, namely two radiologists work independently on reading mammographic images, is the most straightforward and convincing strategy to decrease the proportion of missed cancerous regions or abnormal symptoms [4]. However, it costs a lot of time, labor and energy, and errors are still unavoidable in clinic because of human subjectivity, large variation and fatigue. Fortunately, well-built computer-aided diagnosis (CADx) systems provide radiologists with additional messages to support their image analysis and diagnostic decisions, and release them from this kind of laborious tasks, consequently reducing the death rate in a cost-effective and convenient manner [5].

Massive CADx systems have been explored for specific purposes, such as malignant/benign classification and subtype disease recognition. Most traditional approaches concentrate on handcrafting or selecting efficient and discriminating features, and then integrating one classifier, such as artificial neural network (ANN) [6] and support vector machine (SVM) [7], into the system with tuned parameters by supervised learning. That means careful engineering, tremendous efforts, and professional and empirical knowledge should be paid. Commonly used features are from intensity statistics (kurtosis, area and maximum intensity), shape descriptors (circularity, roughness and elongation), texture descriptors (correlation, contrast and entropy), invariant moments and multi-scale feature descriptors [8, 9]. To distinguish between malignant and benign lesions, Ramos-Pollán *et al*. combined massive features to estimate machine learning classifiers in breast cancer diagnosis [10]. Khan *et al*. investigated the prediction performance of different approaches for directional Gabor feature extraction and weighted SVM [11].

As a branch of machine learning technique, deep learning has recently revolutionized the fields of image representation and object recognition [14]. Different from traditional approaches, deep learning emphasizes on the design of convolutional neural network (CNN) architecture to simulate human-like information abstraction and on the retrieval of low-, mid- and high-level features directly from raw image patches. However, it necessitates large-scale training samples, such as in ImageNet [15], for hyper-parameter tuning and optimization [16]. To tackle the common problem of limited data sets, Arevalo *et al*. designed extremely simple CNN architectures that contain one and two feature layers, respectively named as CNN2 and CNN3, and CNN3 achieves the state-of-the-art performance on a new benchmarking dataset for breast cancer diagnosis [17]. Carneiro *et al*. explored pre-trained CNN model for unregistered multi-view mammogram analysis and achieved promising performance [18], and Benjamin *et al*. investigated learned high-level features as the input of different machine learning classifiers [19].

In this paper, we explore the technique of transfer learning to tackle breast mass classification in mammographic images based on small-scale or limited image samples. Transfer learning is the improvement of learning in a new task through the transfer of knowledge from a related task that has already been well learned and tested [20]. In the field of computer vision, many deep CNN

ICBCB '17, January 06-08, 2017, Hong Kong, Hong Kong

ACM 978-1-4503-4827-0/17/01…$15.00
DOI: http://dx.doi.org/10.1145/3035012.3035022

architectures have been well trained for object reorganization and the models are publicly available. In particular, fine tuning these deep architectures have been successful in medical image analysis and disease diagnosis [21]. In this study, we assume it overcomes the unsatisfactory performance of traditional approaches with handcrafted features and common classifiers, and also break the obstacle of limited data sets for training deep CNNs in clinic.

2. METHODS

2.1 Data sets

BCDR-F03 (Film Mammography dataset number 3) is a newly available dataset on Breast Cancer Digital Repository (BCDR) (http://bcdr.inegi.up.pt) [22]. It is made up of 736 film mammographic images, corresponding to 344 women (mean age = 56±13years, max=98 years, min=21 years) and totally contains 736 biopsy-proven mass lesions (426 benign and 310 malignant mass lesions).

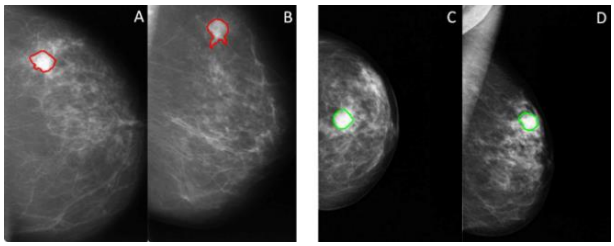

Figure 1. Two representative breast mass lesions

BCDR-F03 can work as a dataset for binary classification (benign vs. malignant findings). Meanwhile, the coordinates of lesions' contours are provided. Figure 1 shows two representative examples. The left ones is a malignant lesion indexed by 68, and the right ones is a benign lesion indexed by 286. They are shown in medio-lateral oblique (MLO) view and craneo-caudal (CC) view, respectively. The accessibility of BCDR-F03 dataset makes possible for novel algorithm design, performance comparison and validation of the research reproducibility in cancer diagnosis.

2.2 Image preprocessing

Image preprocessing aims to enhance tissue characteristics of region of interest (ROI) in the image. In this study, ROIs are first extracted and then Otsu segmentation algorithm is borrowed to roughly isolate the mass region [23]. After that, morphological operation is used to refine the mass. In the end, data augmentation is utilized to increase the sample number.

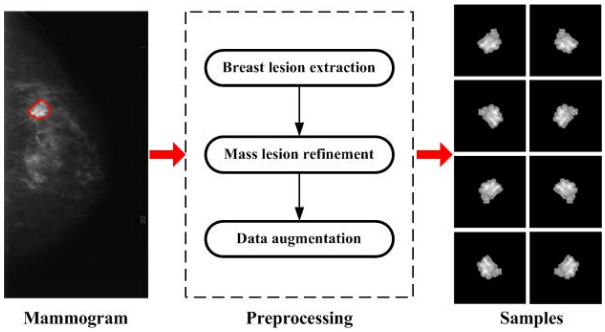

Figure 2. Workflow diagram of data preprocessing

- *Breast lesion extraction*. Suspicious regions provided by radiologists in BCDR-F03 are marked with dotted points in the image. To support the input size of existing AlexNet ([227 277]), a maximum square is generated to crop the mass region and its surrounding tissues. If the square size is larger than 227, we will down-sample it; or else, we will utilize zero padding.
- *Mass lesion refinement*. Extracted mass lesions contain both mass and breast tissues. To isolate the mass lesion regions, Otsu algorithm is used and morphological operation further refines the mass region. Figure 3 shows the results of one example after breast lesion extraction and mass refinement.
- *Data augmentation*. To increase the sample size and the network robustness, data augmentation is used. In breast mass lesion classification, it makes sense because a lesion should be recognized when the lesion is presented in any particular orientation. In this study, to each refined breast lesion, another seven new samples are generated using a combination of flipping and rotation ($\pi/2$, π and $3\pi/2$) transformations.

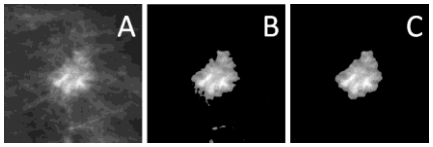

Figure.3. Mass refinement procedure. ROI extraction in (A), Otsu binary segmentation in (B) and refined mass in (C).

2.3 Transfer learning

With a convolutional neural network, its parameters have been pre-trained on ImageNet for classification. Here what we need is to fine-tune the parameters with samples from the dataset BCDR-F03 and transfer these learned generic features on ImageNet for clinical breast cancer diagnosis. The details of the experiment are described as below.

- *Input image size*. Since the refined mass lesion size is [227 227] for AlexNet [24], its size is further cropped to [224 224] for GoogLeNet [25] as in the breast lesion extraction in the image preprocessing.
- *Parameter setting*. In fine tuning, the iteration number is set as 10^4.The primary learning rate is 10^{-3} and it will drop to half once it runs 10^3 iterations. Momentum is 0.9 and weight decay is set 5×10^{-4}. These configurations are to ensure that the parameters are fine tuned for clinical breast cancer diagnosis. Other parameters are set as default values.
- *Software*. In this study, Caffe [26] is utilized under the operation system of Ubuntu 14.04 with 4 Nvidia K80 GPUs. Note that only two GPUs are running.

In the fixed networks of AlexNet and GoogLeNet, the last full connection layer both of which contain 1 thousand class scores is replaced with 2 outputs, indicating benign and malignant mass lesions. In addition, fine tuning is realized with a few samples (114 benign patient cases, and 99 malignant patient cases) to adjust the pre-trained networks for successive approach toward the targeted clinical application by supervised learning and back-propagation. Note that three input channels for CNNs are fed with the same input of breast mass lesion samples.

2.4 Experiment design

Besides pre-trained networks of AlexNet and GoogLeNet, two shallow networks of CNN3 and CNN2, and three algorithms combing various features and SVM are compared. These features include the histogram of gradient divergence (HGD) [27], histogram of oriented gradients (HOG) [28] and gray-level co-occurrence matrix (GLCM) [29]. All algorithm performance is

quantified and compared with values of the area under the receiver operating characteristic curves (AUC).

3. RESULTS

Convolutional kernel visualization gives us an insight into the understanding of learned features, while a comparison of learned features before and after transfer learning allows for describing the patterns that GoogLeNet is looking for in this medical image classification task. Figure 4 demonstrates learned features in the first layer of GoogLeNet after transfer learning, which is fine tuned using breast mass lesion image. Offline subtraction comparison with features learned on ImageNet indicates there is no obvious difference between the features before training and the learned features.

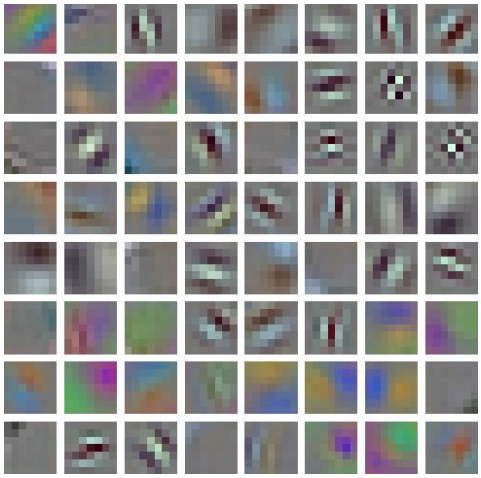

Figure 4.Learned features in the first layer of GoogLeNet.

Table 1 shows quantitative comparison of CNNs. Among these networks, GoogLeNet performs the best (AUC=0.88), followed by AlexNet (AUC=0.83), CNN3 (AUC=0.82), and CNN2 predicts not well. The former two networks are deep networks based on transfer learning, while the latter two are shallow networks that learn from medical image scratches.

Table 1. AUC in the test set from convolutional neural networks and the best result is shown in bold typeface.

	CNN	AUC
Transfer learning	GoogLeNet	**0.88**
Learning from scratch	AlexNet	0.83
	CNN3	0.82
	CNN2	0.76

Table 2 demonstrates AUC values calculated from three traditional approaches. The result indicates that HGD features (AUC=0.78) outperforms other features, followed by HOG (AUC=0.76) and GLCM (AUC=0.68).

Comparing Table 1 and 2, it is observed that GoogLeNet achieves 10 percent increase compared to the best handcrafted features of HGD, while CNN3 is slightly better than traditional methods.

Furthermore, the receiver operating characteristics (ROC) curves from the algorithms based on CNNs are as illustrated in Figure 5. It shows that the overall ability of GoogLeNet is obviously superior over AlexNet, CNN3 and CNN2 to discriminate between benign and malignant breast lesions. It also indicates that CNN2 reaches unsatisfactory performance.

Table 2. AUC in the test set from traditional approaches and the best result is shown in bold type-face.

	Classifier	AUC
HGD	SVM	0.78
HOG	SVM	0.76
GLCM	SVM	0.68

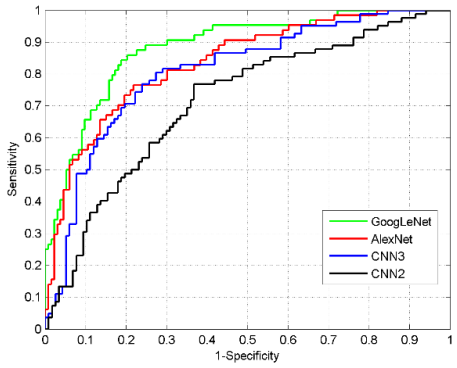

Figure 5. ROC curve for evaluated models

4. DISCUSSION

Well-built CADx systems improve radiologists' diagnostic accuracy when differentiating between benign and malignant masses and reduce unnecessary biopsies in a cost-effective manner. Utilizing professional and empirical knowledge and novel techniques to increase the performance of CADx is always desirable in clinics.

This paper investigated the technique of transfer learning to refresh parameters in deep learning architectures which are pre-trained on large-scale natural images for clinical breast mass lesion classification in mammographic images. Under the context of insufficient data samples in this study, transfer learning that makes use of GoogLeNet for breast cancer diagnosis seems very promising. It enhances the classification performance compared to shallow networks and traditional approaches. Compared with deep CNNs, the capacity of traditional approaches is rather limited and achieves AUC less than 0.8 in prediction. For better prediction performance, combination of various kinds of features is helpful, such as clinical information of age and race, texture analysis and other messages. On the other hand, integrating handcrafted features into features learned from deep networks is also useful. That means, additional low-level handcrafted features to represent the suspicious regions provides useful cues for accurate disease diagnosis. However, feature dimension will increase and thereby leading to over-fitting and low generalization capacity.

Several limitations in this study should be mentioned here. First, the training of deep CNNs might be affected by the proportion of data samples in different classes, i.e. the number between benign and malignant mass lesions. To overcome this defect, training samples can be set equal. Second, the sizes of mass regions are not considered. Further study can take the size into consideration with classifiers for prediction. Third, many CNN networks can be considered, compared and selected, while full comparison is not so practical. In the end, feature extraction

and selection is always a hot field in pattern analysis, and further improvement can be made.

5. CONCLUSION

This paper investigates the technique of transfer learning for breast mass lesion classification which improves prediction performance when compared to shallow neural networks and traditional approaches. Transfer learning not only enhances the unsatisfactory capacity of traditional approaches, but also breaks the obstacle of insufficient data samples for training deeper networks. Until large-scale medical image datasets for specific purposes are available, the combination of transfer learning and deep convolutional neural networks is promising for computer-aided diagnosis.

6. ACKNOWLEDGMENTS

This work is supported partly by Guangdong Innovative Research Team Program (Grant No. 2011S013), CAS Key Laboratory of Human-Machine Intelligence-Synergy Systems, Shenzhen Institutes of Advanced Technology and Shenzhen Fundamental Research Program (JSGG20160229203812944).

7. REFERENCES

[1] R.L. Siegel, K.D. Miller and A. Jemal, Cancer statistics, 2016. CA: a cancer journal for clinicians 66(1) (2016), 7-30.

[2] L. Fan, K. Strasser-Weippl, J.J. Li and *et al*., Breast can-cer in China. The lancet oncology 15(7) (2014), e279-e289.

[3] A. Bleyer and H.G. Welch, Effect of three decades of screening mammography on breast-cancer incidence. New England Journal of Medicine 367(21) (2012), 1998-2005.

[4] J. Brown, S. Bryan and R. Warren, Mammography screening: an incremental cost effectiveness analysis of double versus single reading of mammograms. BMJ 312(7034) (1996), 809-812.

[5] Y. Jiang, R.M. Nishikawa, R.A. Schmidt and *et al*., Improving breast cancer diagnosis with computer-aided diagnosis. Academic radiology, 6(1) (1999), 22-33.

[6] B. Yegnanarayana, Artificial neural networks [M]. PHI Learning Pvt. Ltd., 2009.

[7] C.C. Chang and C.J. Lin, LIBSVM: a library for support vector machines. ACM Transactions on Intelli-gent Systems and Technology (TIST), 2(3)(2011): 27.

[8] K. Ganesan, U.R. Acharya, C.K. Chua and *et al*., Computer-aided breast cancer detection using mammo-grams: a review. IEEE Reviews in Biomedical Engineer-ing 6 (2013), 77-98.

[9] A.K. Jain, R.P.W. Duin and J. Mao. Statistical pat-tern recognition: A review. IEEE Transactions on pattern analysis and machine intelligence 22(1) (2000), 4-37.

[10] R. Ramos-Poll án, M.A. Guevara-L ópez, C. Su árez-Ortega and *et al*., Discovering mammography-based ma-chine learning classifiers for breast cancer diagnosis. Journal of medical systems 36(4) (2012), 2259-2269.

[11] S. Khan, M. Hussain, H. Aboalsamh and *et al*., A comparison of different Gabor feature extraction ap-proaches for mass classification in mammography. Mul-timedia Tools and Applications (2015), 1-25.

[12] Y. LeCun, Y. Bengio and G. Hinton, Deep learning. Nature 521(7553) (2015), 436-444.

[13] J. Deng, W. Dong, R. Socher and *et al*., ImageNet: A large-scale hierarchical image database. CVPR (2009), 248-255.

[14] A. Krizhevsky, I. Sutskever and G.E. Hinton, Imagenet classification with deep convolutional neural networks. NIPS (2012), 1097-1105.

[15] J. Arevalo, F.A. Gonz ález, R. Ramos-Poll án and *et al*., Representation learning for mammography mass le-sion classification with convolutional neural networks. Computer methods and programs in biomedicine 127 (2016), 248-257.

[16] Carneiro G, Nascimento J, Bradley A P. Unregistered multiview mammogram analysis with pre-trained deep learning models[C]//International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer International Publishing, 2015: 652-660.

[17] Huynh B Q, Li H, Giger M L. Digital mammographic tumor classification using transfer learning from deep convolutional neural networks[J]. Journal of Medical Imaging, 2016, 3(3): 034501-034501.

[18] S. J. Pan and Q. Yang, A survey on transfer learn-ing. IEEE Transactions on knowledge and data engineer-ing, 22(10) (2010), 1345-1359.

[19] H. C. Shin, H. R. Roth, M. Gao and *et al*., Deep convolutional neural networks for computer-aided detec-tion: CNN architectures, dataset characteristics and trans-fer learning. IEEE transactions on medical imaging 35(5) (2016), 1285-1298.

[20] L ópez M A G, Posada N G, Moura D C, *et al*. BCDR: a breast cancer digital repository[C]//15th International Conference on Experimental Mechanics. 2012.

[21] N. Otsu, A threshold selection method from gray-level histograms. Automatica, 11(285-296) (1975), 23-27.

[22] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks[C]//Advances in neural information processing systems. 2012: 1097-1105.

[23] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," IEEE CVPR, 1-9 (2015).

[24] Y. Jia, E. Shelhamer, J. Donahue, and *et al*., Caffe: Convolutional architecture for fast feature embedding. Proceedings of the 22nd ACM international conference on Multimedia. ACM, 2014: 675-678.

[25] D.C. Moura and M.A.G. L ópez, An evaluation of image descriptors combined with clinical data for breast cancer diagnosis. International journal of computer assisted radiology and surgery 8(4) (2013), 561-574.

[26] N. Dalal and B. Triggs, Histograms of oriented gradients for human detection. IEEE CVPR, 886-893 (2005).

[27] R.M. Haralick and K. Shanmugam, Textural features for image classification. IEEE Transactions on systems, man, and cybernetics, 1973 (6): 610-621 (1973).