https://doi.org/10.2352/ISSN.2470-1173.2021.10.IPAS-246

Deep Learning Features for Discriminating Between Benign and Malignant Microcalcification Lesions

Juan Wang; Illinois Institute of Technology; Chicago, IL, USA
Liang Lei; Chongqing University of Science and Technology; Chongqing, China
Yongyi Yang; Illinois Institute of Technology; Chicago, IL, USA

Abstract

Accurate diagnosis of microcalcification (MC) lesions in mammograms as benign or malignant is a challenging clinical task. In this study we investigate the potential discriminative power of deep learning features in MC lesion diagnosis. We consider two types of deep learning networks, of which one is a convolutional neural network developed for MC detection and the other is a denoising autoencoder network. In the experiments, we evaluated both the separability between malignant and benign lesions and the classification performance of image features from these two networks using Fisher's linear discriminant analysis on a set of mammographic images. The results demonstrate that the deep learning features from the MC detection network are most discriminative for classification of MC lesions when compared to both features from the autoencoder network and traditional handcrafted texture features.

Introduction

Breast cancer is the most frequently diagnosed non-skin cancer in women in the US [1]. The occurrence of clustered microcalcifications (MCs) can be an important early sign of breast cancer in mammograms. MCs are tiny calcium deposits that appear as small spots (typically 0.1–1 mm in diameter) in mammogram images (Fig. 1). They can be found in both malignant and benign cases in mammogram screening. Due to the subtlety of MCs in mammograms, accurate diagnosis of MC lesions as benign or malignant is a very challenging clinical task [2].

Because of the difficulty in diagnosis of MC lesions, there have been great efforts in the literature in developing computer-aided diagnosis (CAD) methods for discriminating between malignant and benign MC lesions [3-10]. One such effort is on development of various machine learning algorithms for MC lesion classification. For example, several pattern classifier models were studied in [3] for differentiating malignant from benign MC lesions. An adaptive Adaboost classifier was developed in [4] to improve the performance of MC lesion diagnosis by retrieving similar images from a library of known cases. A two-step classification approach was proposed in [5] based on view-level decisions for MC lesion diagnosis. A deep learning approach was developed in [6] to classify malignant and benign MC lesions.

Parallel to the development of classifier algorithms in CAD, there are also great interests in development of discriminative image features for characterizing MC lesions. These features were typically handcrafted based on the image properties (e.g., size, shape, etc.) of the MC objects in a lesion. For example, in [7] a set of image features was developed based on radiologists' interpretation of malignant and benign MC lesions. In [8], texture features were studied for MC lesion diagnosis. In [9], a set of quantitative features based on spatial modeling of clustered MCs was used to describe MC lesions. In [10], graph theoretical features were developed for MC lesion classification.

In recent years, deep convolutional neural networks (CNNs) are demonstrated to yield image features that can be more discriminative than hand-crafted features in many pattern classification tasks in image processing [11, 12]. For the CAD task of MC lesion diagnosis, the extraction of hand-crafted image features from a given MC lesion typically requires the knowledge of the locations of the individual MCs, which itself is a challenging task that is either time consuming when marked manually or subject to false positives (or missed detections) when detected by a computerized algorithm. Therefore, as an alternative, using deep learning features in CAD can potentially avoid the above difficulty in characterizing MC lesions by using handcrafted features.

In our recent study [13], we investigated the use of image features derived from deep learning models for modeling perceptually similar MC lesions, and found that they can yield a good agreement with the perpetual similarity between MC lesions as judged by radiologists. Encouraged by this result, in this study we further investigate whether deep learning features can also be discriminative for malignant and benign MC lesion classification. For this purpose, we consider two types of deep learning features, of which one is derived from a supervised-learning network for the task of MC detection, and the other is from an unsupervised-learning autoencoder network. For the former, the global MC cluster detector network developed in [14] is used. For the latter, a denoising autoencoder network [15] is used. To quantify the discriminative power of the deep learning features, we applied Fisher's linear discriminant analysis on a set of malignant and benign lesion images in the experiments. For comparison, we also considered a set of commonly used texture features for MC lesions [16].

Methods

Motivation

As noted in the introduction, the goal of this study is to investigate whether the image features obtained from deep learning networks trained for MC detection can be discriminative between malignant and benign MC lesions. For this purpose, we consider two deep learning networks, of which one is trained for the task of MC cluster detection and the other is an autoencoder network trained for image representation. Below we describe the details of these two networks.

Supervised deep learning features

We first consider a CNN network previously developed for detecting the presence of MC clusters in mammograms [14]. This network was trained to discriminate image regions containing clustered MCs from those without any MCs. Thus, the feature maps generated from this network are expected to capture the

important image features relevant to MC lesions, as previously illustrated in [14].

The MC detector network in [14] consisted of five convolutional (Conv), ReLU and local response normalization (LRN) layers, four max-pooling layers (denoted as p1-p4), and two fully-connected layers. For this study, this detector model is retrained with the following minor modifications: 1) the LRN layers are removed, and 2) the combination of Conv and ReLU layers is replaced by a combination of Conv, batch normalization (BN) and ReLU layers. Such modifications are to avoid the need for tuning the hyper-parameters associated with the LRN layers.

To extract the CNN features for an MC lesion, the image region of the lesion is first fed to the network. Afterward, the feature maps are obtained from the different layers of the network. In this study, the feature maps from the four max-pooling layers p1-p4 are used, which represent features extracted at different scales. In order to reduce the number of features, a global max-pooling is applied to each feature map as in [17]. In the end, this yields 32, 64, 128, 128 features for layers p1-p4, respectively. These features are used to form a vector **x** (i.e., feature vector) for representing the lesion region.

Unsupervised deep learning features

We next consider a denoising autoencoder network, which is commonly used for efficient data representation [15]. Conceptually, an autoencoder network is trained to reproduce an input signal at the output with a representation (i.e., feature maps) of a much lower dimension. Hence, it is selected for study here on whether it can effectively capture the important diagnostic image features of MC lesions.

Specifically, the autoencoder network used in this study is formed by three Conv+BN+ReLU blocks and three max-pooling layers (denoted as p1-p3) in the encoder; correspondingly, the decoder is formed by two Conv+BN+ReLU blocks, one Conv+BN+Sigmoid block, and three up-sampling layers. Here Sigmoid denotes a sigmoid activation layer, which produces output within [0, 1]. In the Conv layers, the kernel size and number of kernels are all set as 3×3 and 64, respectively. In the max-pooling layers, the stride and pooling size are all set as 2 and 2×2, respectively.

To extract the encoder features, as in the MC detector network above, the feature maps from the three max-pooling layers p1-p3 are used. Also, a global average-pooling is applied to each feature map to reduce the number of features, yielding a total of 64 features for each layer.

Handcrafted texture features

For comparison, we also consider a set of textural measures which have been widely used for characterizing MC lesion regions in mammograms [16]. These features are derived from the spatial gray level dependence (SGLD) matrices of a lesion region, and do not need the locations of individual MCs. Specifically, the following 12 features are extracted for each lesion [16]: energy, entropy, difference average, difference variance, difference entropy, sum average, sum variance, sum entropy, inverse difference moment, correlation, and two information measures of correlation.

Fisher's linear discriminant analysis

To examine the potential discriminative power of deep learning features in MC lesion diagnosis, we apply Fisher's linear discriminant analysis on separating between malignant and benign MC lesions. We first assess the separability between the two classes by the deep learning features, then quantify the classification performance of the features by using the Fisher linear discriminant on a set of MC lesions.

Fisher's linear discriminant is a linear classifier which projects a high-dimensional feature vector **x** onto a hyperplane that separates the feature space into two half-spaces (corresponding to two classes). The classifier function is given by

$$y = \mathbf{w}^T\mathbf{x} + b \tag{1}$$

where **w** is the Fisher discriminant vector, and b is the decision bias. Mathematically, the discriminate vector **w** is obtained by maximizing the Fisher separation criterion [9], which is given by the ratio of between-class variance to within-class variance as

$$J = \frac{\mathbf{w}^T S_b \mathbf{w}}{\mathbf{w}^T S_w \mathbf{w}} \tag{2}$$

where S_b and S_w denote the between-class and within-class scatter matrices, respectively, of the two classes. The resulting Fisher discriminant vector is

$$\mathbf{w} = S_w^{-1}(\boldsymbol{m}_1 - \boldsymbol{m}_0) \tag{3}$$

where $\boldsymbol{m}_1$ and $\boldsymbol{m}_0$ are the respective mean feature vectors of the two classes.

Note that a larger value in the Fisher separation criterion J corresponds to a better separation by the classifier between the data samples from the two classes, and vice versa. Thus, in the preliminary study below, the class separation criterion J is used to quantify the potential discriminative power of the image features derived from the deep learning networks.

Furthermore, to evaluate the classification performance of the deep learning features in MC lesion diagnosis, we apply the Fisher's linear classifier using a 5-fold cross validation procedure on a set of MC lesions. In the experiments, the set of cases was randomly divided into five equally-sized subsets, then the classifier was trained and tested by using each of the five subsets in turn as test set while the rest as training set. To quantify the classification performance, we conduct a receiver-operating characteristic (ROC) analysis, which is now routinely used for performance evaluation in a binary classification task. An ROC curve is a plot of the true-positive rate versus the false-positive rate when the operating threshold is varied continuously. The area under the ROC curve, denoted by AUC, is used to summarize the diagnostic performance. A larger AUC value means better classification performance.

Given the high dimension of the deep learning features in the two networks, in our experiments a principle component analysis (PCA) was first applied to the deep learning features in which the first 12 most dominant components were used in Fisher's linear discriminant analysis. This preprocessing step was to accommodate the limited number of data samples and for fair comparison with the texture features.

Experiments

Dataset for performance evaluation

To evaluate the discriminative performance of the deep learning features in MC lesion diagnosis, we made use of a dataset collected by the Department of Radiology at the University of Chicago. It consists of 408 mammogram images, all containing

lesions with MCs, from 238 cases (118 malignant cases and 120 benign cases). All of these cases had been pathologically proven by biopsy, which are used as ground truth in our performance evaluation. For each mammogram, regions of interest (ROIs) were cropped in size of 512×512 or 1,024×1,024 pixels with a spatial resolution of 0.1 mm/pixel based on the size of the lesions.

MC detector and autoencoder network training

For training the MC detector and autoencoder, we made use of an independent set of mammogram images (no overlapping cases with the evaluation dataset above). It consisted of 113 screen-film and 188 full-field digital mammogram images, all with spatial resolution of 0.1 mm/pixel, collected by the Department of Radiology at the University of Chicago. Each image had at least one cluster of MCs. All the images were pre-processed for tissue background suppression [14]. These images were partitioned randomly into two non-overlapping subsets as follows: 1) a subset with 241 images for network training, and 2) a subset with 60 images for model validation.

For training the MC detector network, image patches of 95×95 were extracted from the training set of mammogram images as in [14], and used as training samples (with or without MCs). Each image patch was normalized to have zero mean and unit standard deviation for input to the network.

For training the denoising autoencoder network, image patches of 96×96 pixels were extracted from the MC regions of the training mammograms. Each image patch was scaled within the range of [0, 1]; afterward, Gaussian noise (mean 0 and standard deviation 0.1) was added. The resulting noisy image patch was then clipped in value within [0, 1], and was applied as input to the network. The network was trained to reproduce the image patch without any added noise.

For both networks, a binary cross-entropy loss was used during training. The adaptive moment estimation (Adam) method was used for optimization. Our implementation was based on the Keras package with Tensorflow backend.

Results and Discussions

Class separability by deep learning features

To demonstrate the class separability in MC lesion diagnosis by deep learning features, we show in Table 1 the Fisher separation criterion results for the features extracted at different scale levels of the two deep learning networks using all the cases in the evaluation dataset. Specifically, for the MC detector network, the Fisher separation criterion is given for each of the four max-pooling layers p1-p4 of the network. Similarly, for the denoising autoencoder network, the Fisher separation criterion results are given for the three max-pooling layers p1-p3 of the network. In addition, for comparison, the Fisher separation criterion was also computed for the texture features to be 0.308.

As can be seen, for the MC detector network the features from the third pooling layer p3 achieved the best separation value of 0.355. Interestingly, it is observed in Table 1 that the Fisher separation criterion value shows an increasing trend for the first three pooling layers p1-p3 in the MC detector network. These results show that the intermediate level features from the MC detector network are most discriminative for MC lesion diagnosis. This may indicate that these features can better characterize the image properties of individual MCs and their spatial clustering properties.

For the autoencoder network, the features from the second pooling layer p2 achieved the best correlation value 0.195, which is notably lower than that from the texture features.

Classification performance by deep learning features

To demonstrate the performance of deep learning features in classification of malignant and benign MC lesion, we show in Table 2 the AUC values obtained by the Fisher linear classifier using the deep learning features extracted at different scale levels for both the MC detector network and denoising autoencoder network. For comparison, the AUC value was also obtained for the texture features as 0.648.

As can be seen, for the MC detector network the features from the third pooling layer p3 achieved the highest AUC value of 0.669. Moreover, similar to the results earlier in Table 2, the AUC value also shows an increasing trend for features among the first three pooling layers in the MC detector network.

For the autoencoder network, the features from the second pooling layer p2 achieved the highest AUC value of 0.561, which is lower than that from the texture features.

Discussions

It is noted from the results in Tables 1 and 2 that, while the deep learning features from the MC detector network are demonstrated to have discriminative power for MC lesion diagnosis, the achieved AUC values are relatively low. We believe that this underlines the difficulty of the cases under consideration. Indeed, these low AUC values were consistent with several observer studies reported on MC lesion diagnosis in the literature. For example, an average AUC value of 0.61 was obtained in an observer study conducted with 10 radiologists in [2], and AUC values in the range of 0.61 to 0.79 were reported in an observer study for a group of 12 breast radiologists in [18].

It is also noted in Tables 1 and 2 that the discriminative power can vary greatly for different deep learning features. The features from the denoising autoencoder network are found to be far less discriminative than those from the MC detector network. We believe that a possible reason for this difference is the following: Because MCs are tiny objects in mammograms (Fig. 1), they cannot be well represented by a general-purpose denoising autoencoder network; in contrast, the MC detector network was trained to identify the presence of MC clusters in a mammogram image, thus better able to capture the image characteristics of individual MCs and MC clusters.

As a feasibility study, we considered only the Fisher linear classifier for MC lesion diagnosis in the experiments. In the future, it would be interesting to further investigate the use of nonlinear classifiers, which are expected to yield improved classification performance. In addition, the MC detector network was trained for the relevant task of MC detection on a set of MC lesions (in which there are many individual MCs); it would also be interesting to directly train a deep learning network for MC lesion diagnosis, though such an approach would require a much larger number of available cases (with known diagnosis) in order to achieve optimal generalizability.

Conclusion

We investigated the potential discriminative power of deep learning features in MC lesion diagnosis. Two types of deep neural networks were considered in the experiments, of which one is a CNN detector trained for MC detection, and the other is a denoising autoencoder network for image representation. The features extracted from the MC detector network were found to be more powerful than those extracted from the denoising autoencoder network. The features from the MC detector network are also more discriminative than the handcrafted texture features between malignant and benign MC lesions. In the future, it would be interesting to investigate whether using nonlinear classifiers on deep learning features can further improve the classification performance of malignant and benign MC lesions.

References

[1] American Cancer Society, "Cancer facts and figures," Atlanta, GA, 2019.

[2] Y. Jiang, R. M. Nishikawa, R. A. Schmidt, C. E. Metz, M. L. Giger, and K. Doi, "Improving breast cancer diagnosis with computer-aided diagnosis," *Academic Radiology*, vol. 6, no. 1, pp. 22-33, 1999.

[3] L. Wei, Y. Yang, R. M. Nishikawa, and Y. Jiang, "A study on several machine-learning methods for classification of malignant and benign clustered microcalcifications," *IEEE Transactions on Medical Imaging*, vol. 24, no. 3, pp. 371-380, 2005.

[4] J. Wang and Y. Yang, "Boosted classification of breast cancer by retrieval of cases having similar disease likelihood," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 908-911, 2016.

[5] A. J. Bekker, M. Shalhon, H. Greenspan, and J. Goldberger, "Multi-view probabilistic classification of breast microcalcifications," *IEEE Transactions on Medical Imaging*, vol. 35, no. 2, pp. 645-653, 2015.

[6] J. Wang, X. Yang, H. Cai, W. Tan, C. Jin, and L. Li, "Discrimination of breast cancer with microcalcifications on mammography by deep learning," *Scientific Reports*, vol. 6, no. 1, pp. 1-9, 2016.

[7] Y. Jiang, R. M. Nishikawa, D. E. Wolverton, C. E. Metz, M. L. Giger, R. A. Schmidt, C. J. Vyborny, and K. Doi, "Malignant and benign clustered microcalcifications: automated feature analysis and classification," *Radiology*, vol. 198, no. 3, pp. 671-678, 1996.

[8] A. N. Karahaliou, I. S. Boniatis, S. G. Skiadopoulos, F. N. Sakellaropoulos, N. S. Arikidis, E. A. Likaki, G. S. Panayiotakis, and L. I. Costaridou, "Breast cancer diagnosis: analyzing texture of tissue surrounding microcalcifications," *IEEE Transactions on Information Technology in Biomedicine*, vol. 12, no. 6, pp. 731-738, 2008.

[9] J. Wang and Y. Yang, "Spatial density modeling for discirminating between benign and malignant microcalcification lesions," *IEEE International Symposium on Biomedical Imaging (ISBI)*, pp. 133-136, 2013.

[10] Z. Chen, H. Strange, A. Oliver, E. RE Denton, C. Boggis, and R. Zwiggelaar, "Topological modeling and classification of mammographic microcalcification clusters," *IEEE Transactions on Biomedical Engineering*, vol. 62, no. 4, pp. 1203-1214, 2014.

[11] G. Antipov, S. Berrani, N. Ruchaud, and J. Dugelay, "Learned vs. hand-crafted features for pedestrian gender recognition," *ACM International Conference on Multimedia*, pp. 1263-1266. 2015.

[12] A. S. Keçeli, A. Kaya, and S. U. Keçeli, "Classification of radiolarian images with hand-crafted and deep features," *Computers & Geosciences*, vol. 109, 67-74, 2017.

[13] J. Wang, L. Lei, and Y. Yang, "Deep learning features for modeling perceptual similarity in microcalcification lesion retrieval," *IEEE International Symposium on Biomedical Imaging (ISBI)*, 2020.

[14] J. Wang, R. M. Nishikawa, and Y. Yang, "Global detection approach for clustered microcalcifications in mammograms using a deep learning network," *Journal of Medical Imaging*, vol. 4, no. 2, pp. 024501, 2017.

[15] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol, "Extracting and composing robust features with denoising autoencoders," *ACM International Conference on Machine Learning*, pp. 1096-1103, 2008.

[16] J. Wang, R. M. Nishikawa, and Y. Yang, "Quantitative comparison of clustered microcalcifications in for-presentation and for-processing mammograms in full-field digital mammography," *Medical Physicals*, vol. 44, no. 7, pp. 3726-3738, 2017.

[17] M. Chen, S. Shi, Y. Zhang, D. Wu, and M. Guizani, "Deep features learning for medical image analysis with convolutional autoencoder neural network," *IEEE Transactions on Big Data*, 2017.

[18] J. Wang, Y. Yang, M. N. Wernick, and R. M. Nishikawa, "An image-retrieval aided diagnosis system for clustered microcalcifications," *IEEE International Symposium on Biomedical Imaging (ISBI)*, pp. 1076-1079, 2016.

Author Biography

Juan Wang received her BS and MS degrees from University of Electronic Science and Technology of China in 2007 and 2010, respectively, and her PhD degree from Illinois Institute of Technology in 2015. She is an Information Scientist at the Delta Micro Technology Inc. Her work focuses on computer-aided diagnosis, medical imaging, machine learning, and deep learning.

Liang Lei is currently a professor with the Department of Computer Science, Chongqing University of Science and Technology, Chongqing, China. His research interests lie in the areas of information retrieval and computer network security.

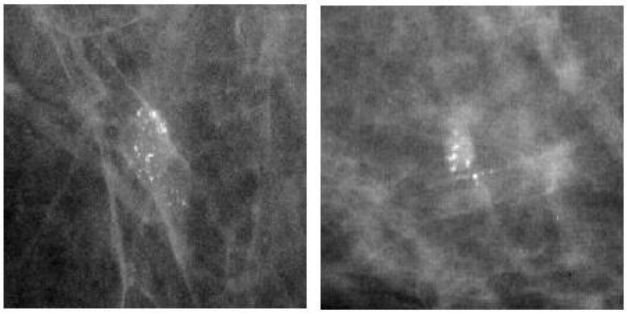
Yongyi Yang is currently a Harris Perlstein Professor with the Department of Electrical and Computer Engineering, Illinois Institute of Technology. His recent research activities are mostly in computerized techniques for breast cancer detection and diagnosis, and in image reconstruction methods for cardiac diagnostic imaging. His research interests include medical imaging, machine learning, pattern recognition, and biomedical applications. He has authored or coauthored over 250 peer-reviewed publications in these areas. He is a fellow of the American Institute for Medical and Biological Engineering (AIMBE).

Table 1. Fisher's seperation criterion for different features in discriminating between malignant and benign MC lesions.

Network	p1	p2	p3	p4
MC detector	0.277	0.299	**0.355**	0.267
Autoencoder	0.149	**0.195**	0.066	-

Table 2. AUC values of different features in discriminating between malignant and benign MC lesions.

Network	p1	p2	p3	p4
MC detector	0.658	0.652	**0.669**	0.646
Autoencoder	0.520	**0.561**	0.520	-



(a) (b)

Figure 1. Examples of mammogram ROIs containing clustered MCs: (a) a malignant lesion, and (b) a benign lesion. The MCs are tiny calcium deposits which appear as small white spots. The image contrast has been adjusted in these ROIs in order to enhance the visibility of the MCs.

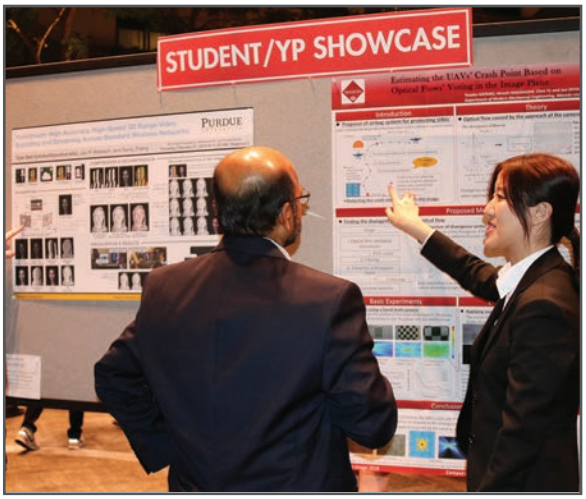
JOIN US AT THE NEXT EI!

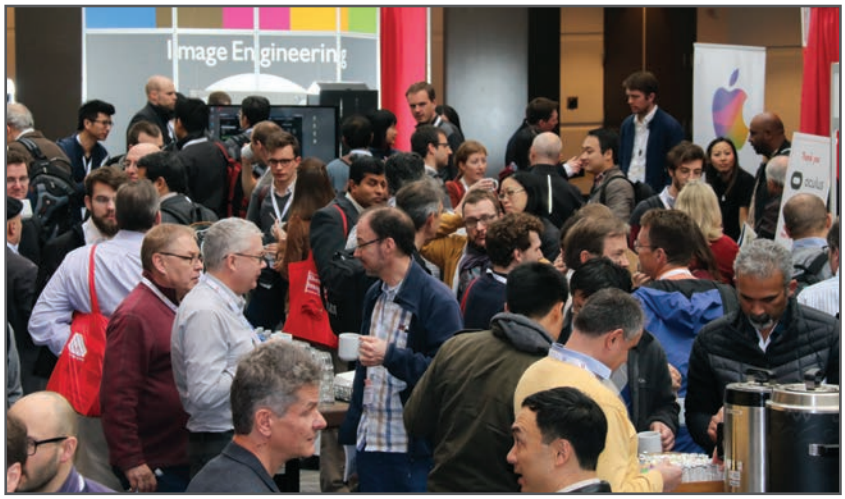
IS&T International Symposium on

Electronic Imaging

SCIENCE AND TECHNOLOGY

Imaging across applications . . . Where industry and academia meet!





• SHORT COURSES • EXHIBITS • DEMONSTRATION SESSION • PLENARY TALKS •
• INTERACTIVE PAPER SESSION • SPECIAL EVENTS • TECHNICAL SESSIONS •

www.electronicimaging.org

IS&T
imaging.org