Deep Multi-instance Networks with Sparse Label Assignment for Whole Mammogram Classification

Wentao Zhu, Qi Lou, Yeeleng Scott Vang, and Xiaohui Xie

{wentaoz1, xhx}@ics.uci.edu, {qlou, ysvang}@uci.edu

Abstract. Mammogram classification is directly related to computer-aided diagnosis of breast cancer. Traditional methods requires great effort to annotate the training data by costly manual labeling and specialized computational models to detect these annotations during test. Inspired by the success of using deep convolutional features for natural image analysis and multi-instance learning for labeling a set of instances/patches, we propose end-to-end trained deep multi-instance networks for mass classification based on whole mammogram without the aforementioned costly need to annotate the training data. We explore three different schemes to construct deep multi-instance networks for whole mammogram classification. Experimental results on the INbreast dataset demonstrate the robustness of proposed deep networks compared to previous work using segmentation and detection annotations in the training.

Keywords: Deep multi-instance learning, whole mammogram classification, max pooling-based multi-instance learning, label assignment-based multi-instance learning, sparse multi-instance learning

1 Introduction

According to the American Cancer Society, breast cancer is the most frequently diagnosed solid cancer and the second leading cause of cancer death among U.S. women. Mammogram screening has been demonstrated to be an effective way for early detection and diagnosis, which can significantly decrease breast cancer mortality [17]. However, screenings are usually associated with high false positive rates, high variability among different clinicians, and over-diagnosis of insignificant lesions [17]. To address these issues, it is important to develop fully automated robust mammographic image analysis tools that can increase detection rate and meanwhile reduce false positives.

Traditional mammogram classification requires extra annotations such as bounding box for detection or mask ground truth for segmentation. These methods rely on hand-crafted features from mass region followed by classifiers [20]. The main barrier to use hand-crafted features is the associated cost of time and effort. Besides, these features have potential poor transferability for use in other problem settings because they are not data driven. Other works have employed different deep networks to detect region of interest (ROI) and obtained mass boundaries in different stages [7]. However, these methods require training data to be annotated with bounding boxes and segmentation ground truths which require expert domain knowledge and costly effort to obtain.

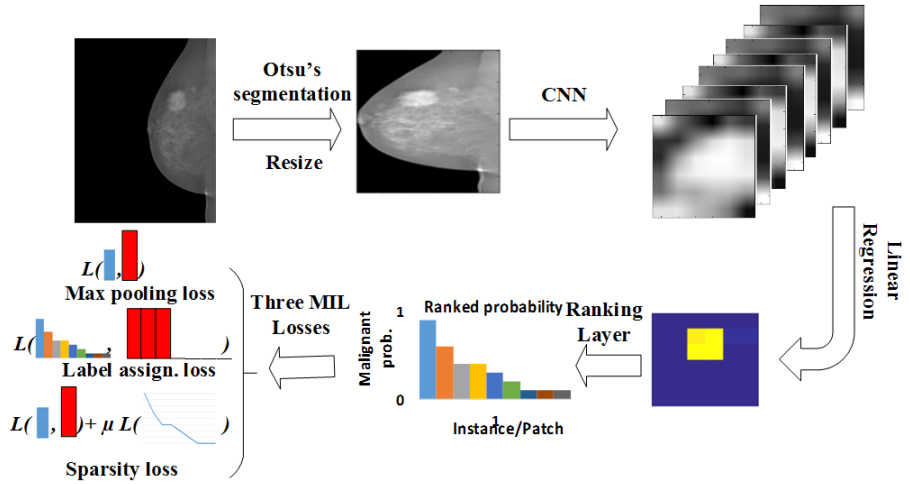
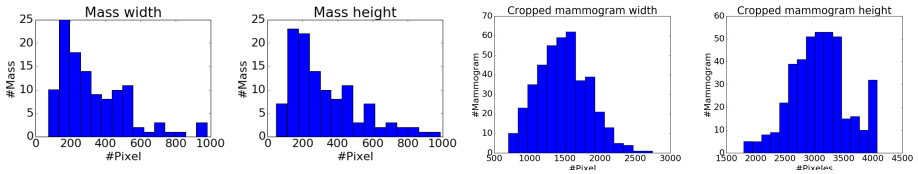


Fig. 1. The proposed deep multi-instance network framework. First, we use Otsu's segmentation to remove the background and resize the mammogram to 224×224. Second, the deep multi-instance network accepts the resized mammogram as input to the convolutional layers. Third, linear regression with weight sharing is employed for the malignant probability of each position from the convolutional neural network (CNN) feature maps of high channel dimensions. Then the responses of the instances/patches are ranked. Lastly, the learning loss is calculated using max pooling loss, label assignment, or sparsity loss for the three different schemes.

Due to the high cost of annotation, we intend to perform classification based on a raw, un-annotated whole mammogram. Each patch of a mammogram can be treated as an instance and a whole mammogram is treated as a bag of instances. The whole mammogram classification problem can then be thought of as a standard multi-instance learning problem. Thus, we propose three different schemes, i.e., max pooling, label assignment, and sparsity, to perform deep multi-instance learning for the whole mammogram classification task.

The framework for our proposed end-to-end deep multi-instance networks for mammogram classification is shown in Fig. 1. To fully explore the power of deep multi-instance network, we convert the traditional multi-instance learning assumption into a label assignment problem. Specifically, we also propose a more efficient, label assignment based deep multi-instance network. As a mass typically composes only 2% of a whole mammogram (see Fig. 2), we further propose sparse deep multi-instance network which is a compromise between max pooling-based and label assignment-based multi-instance networks. The proposed deep multi-instance networks are shown to provide robust performance for whole mammogram classification on the INbreast dataset [16].



(a) (b) (c) (d)

Fig. 2. Histograms of mass width (a) and height (b), mammogram width (c) and height (d). Compared to the size of whole mammogram ($1,474 \times 3,086$ on average after cropping), the mass of average size (329×325) is tiny, and takes about 2% of a whole mammogram.

2 Related Work

2.1 Mammogram Classification

Beura et al. designed co-occurrence features and used wavelet transform for breast cancer detection [4]. Several works have used deep networks to perform mammogram mass classification [11,6,25]. However, those methodologies require annotated mass ROI and/or segmentation ground truth. Dhungel et al. trained a detector and segmentation network on the training set first, and then used CNN to perform mass classification [7]. The training procedure still requires detection ROI and boundary ground truth, which is costly. In addition, multi-stage training cannot fully explore the power of the deep network. Thus, an end-to-end approach for whole mammogram classification is preferred for this problem.

2.2 Deep Multi-instance Learning

Dietterich et al. first proposed multi-instance learning problem [8]. There are various other multi-instance related work in the machine learning literature. Andrews et al. generalized support vector machine for the multi-instance problem [1]. Kwok and Cheung employed marginalized kernel to solve the instance label ambiguity in multi-instance learning [14]. Zhou et al. extended multi-instance learning to multi-class classification problems [23].

Due to the great representation power of deep features [24], combining multi-instance learning with deep neural networks is an emerging topic. Wu et al. combined CNN with multi-instance learning to auto-annotate natural images [21]. Kotzias et al. incorporated CNN features into multi-instance cost function to do sentiment analysis [12]. Yan et al. used a deep multi-instance network to find discriminative patches for body part recognition [22]. Patch based CNN added a new layer after the last layer of deep multi-instance network to learn the fusion model for multi-instance predictions [10]. The above approaches used max pooling to model the general multi-instance assumption which only considered the patch of max probability. In this paper, a more effective task-related deep multi-instance models are explored for whole mammogram classification.

3 Deep Multi-instance Networks for Whole Mammogram Mass Classification

Leveraging the insights from recent successful deep convolution networks used for natural image processing, we design end-to-end trained deep multi-instance networks for the task. Fig. 1 shows the proposed network architecture which has multiple convolutional layers, one linear regression layer, one ranking layer, and one multi-instance loss layer. We employ three schemes for combining multiply instances, 1) the max pooling-based multi-instance learning takes only the largest element from the ranking layer; 2) label assignment-based multi-instance learning utilizes all the elements; and 3) sparse multi-instance learning adds sparse constraints for elements to the ranking layer. The details of these schemes will be detailed later.

The rest of this section is organized as follows. We first briefly introduce the common part of the deep multi-instance networks to make the paper self-contained. Then we introduce the max pooling-based deep multi-instance network in section 3.1. After that, we convert the multi-instance learning into a label assignment problem in section 3.2. Lastly section 3.3 describes how to inject the priori knowledge that a mass comprises small percentage of a whole mammogram into the deep multi-instance network.

CNN is a successful model to extract deep features from images [15]. Unlike other deep multi-instance network [22,10], we use a CNN to efficiently obtain features of all patches (instances) at the same time. Given an image $\boldsymbol{I}$, we can get a much smaller feature map $\boldsymbol{F}$ of multi-channels N_c after multiple convolutional layers and max pooling layers. The $(\boldsymbol{F})_{i,j,:}$ represents deep CNN features for a patch $\boldsymbol{Q}_{i,j}$ in $\boldsymbol{I}$, where i, j represents the pixel row and column indices respectively, and : denotes the channel dimension.

The goal of our work is to predict whether a whole mammogram contains a malignant mass (BI-RADS $\in \{4, 5, 6\}$ as positive) or not, which is a standard binary class classification problem. We add a logistic regression with weights shared across all the pixel positions following $\boldsymbol{F}$. After that, an element-wise sigmoid activation function is applied to the output. The malignant probability of feature space's pixel (i, j) is

$$r_{i,j} = \text{sigmoid}(\boldsymbol{a} \cdot \boldsymbol{F}_{i,j,:} + b), \tag{1}$$

where $\boldsymbol{a}$ is the weights in logistic regression, and b is the bias, and $\cdot$ is the inner product of the two vectors $\boldsymbol{a}$ and $\boldsymbol{F}_{i,j,:}$. The $\boldsymbol{a}$ and b are shared for different pixel position i, j. We can combine $r_{i,j}$ into a matrix $\boldsymbol{r} = (r_{i,j})$ of range $[0, 1]$ denoting the probabilities of patches being malignant masses. The $\boldsymbol{r}$ can be flattened into a one-dimensional vector as $\boldsymbol{r} = (r_1, r_2, ..., r_m)$ corresponding to flattened patches $(\boldsymbol{Q}_1, \boldsymbol{Q}_2, ..., \boldsymbol{Q}_m)$, where m is the number of patches.

3.1 Max Pooling-based Multi-instance Learning

The general multi-instance assumption is that if there exists an instance that is positive, the bag is positive. The bag is negative if and only if all instances are negative [8]. For whole mammogram classification, the equivalent scenario is that if there exists a malignant mass, the mammogram $\boldsymbol{I}$ should be classified as positive. Likewise, negative

mammogram $\boldsymbol{I}$ should not have any malignant masses. If we treat each patch $\boldsymbol{Q}_i$ of $\boldsymbol{I}$ as an instance, the whole mammogram classification is a standard multi-instance task.

For negative mammograms, we expect all the r_i to be close to 0. For positive mammograms, at least one r_i should be close to 1. Thus, it is natural to use the maximum component of $\boldsymbol{r}$ as the malignant probability of the mammogram $\boldsymbol{I}$

$$p(y=1|\boldsymbol{I},\boldsymbol{\theta}) = \max\{r_1, r_2, ..., r_m\}, \tag{2}$$

where $\boldsymbol{\theta}$ is the parameters of deep networks.

If we sort $\boldsymbol{r}$ first in descending order as illustrated in Fig. 1, the malignant probability of the whole mammogram $\boldsymbol{I}$ is the first element of ranked $\boldsymbol{r}$ as

$$\begin{aligned} &\{r'_1, r'_2, ..., r'_m\} = \text{sort}(\{r_1, r_2, ..., r_m\}), \\ &p(y=1|\boldsymbol{I},\boldsymbol{\theta}) = r'_1, \quad \text{and} \quad p(y=0|\boldsymbol{I},\boldsymbol{\theta}) = 1 - r'_1, \end{aligned} \tag{3}$$

where $\boldsymbol{r}' = (r'_1, r'_2, ..., r'_m)$ is descending ranked $\boldsymbol{r}$. The cross entropy-based cost function can be defined as

$$\mathcal{L}_{maxpooling} = -\sum_{n=1}^{N} \log(p(y_n|\boldsymbol{I}_n,\boldsymbol{\theta})) + \frac{\lambda}{2}\|\boldsymbol{\theta}\|^2 \tag{4}$$

where N is the total number of mammograms, $y_n \in \{0, 1\}$ is the true label of malignancy for mammogram $\boldsymbol{I}_n$, and λ is the regularizer that controls model complexity.

Typically, a mammogram dataset is imbalanced, (e.g., the proportion of positive mammograms is about 20% for the INbreast dataset). In lieu of that, we introduce a weighted loss defined as

$$\mathcal{L}_{maxpooling} = -\sum_{n=1}^{N} w_{y_n} \log(p(y_n|\boldsymbol{I}_n,\boldsymbol{\theta})) + \frac{\lambda}{2}\|\boldsymbol{\theta}\|^2, \tag{5}$$

where w_{y_n} is the empirical estimation of y_n on the training data.

One disadvantage of max pooling-based multi-instance learning is that it only considers the patch $\boldsymbol{Q}'_1$ (patch of the max malignant probability), and does not exploit information from other patches. A more powerful framework should add task-related priori, such as sparsity of mass in whole mammogram, into the general multi-instance assumption and explore more patches for training.

3.2 Label Assignment-based Multi-instance Learning

For the conventional classification tasks, we assign a label to each data point. In the multi-instance learning scheme, if we consider each instance (patch) $\boldsymbol{Q}_i$ as a data point for classification, we can convert the multi-instance learning problem into a label assignment problem.

After we rank the malignant probabilities $\boldsymbol{r} = (r_1, r_2, ..., r_m)$ for all the instances (patches) in a whole mammogram $\boldsymbol{I}$ using the first equation in Eq. 3, the first few r'_i should be consistent with the label of whole mammogram as previously mentioned,

while the remaining patches (instances) should be negative. Instead of adopting the general multi-instance learning assumption that only considers the $\boldsymbol{Q'}_1$ (patch of malignant probability r'_1), we assume that 1) patches of the first k largest malignant probabilities $\{r'_1, r'_2, ..., r'_k\}$ should be assigned with the same class label as that of whole mammogram, and 2) the rest patches should be labeled as negative in the label assignment-based multi-instance learning.

After the ranking layer using the first equation in Eq. 3, we can obtain the malignant probability for each patch

$$p(y=1|\boldsymbol{Q'}_i, \boldsymbol{\theta}) = r'_i, \quad \text{and} \quad p(y=0|\boldsymbol{Q'}_i, \boldsymbol{\theta}) = 1 - r'_i. \tag{6}$$

The weighted cross entropy-based loss function of the label assignment-based multi-instance learning can be defined as

$$\begin{aligned} \mathcal{L}_{labelassign.} = -\sum_{n=1}^{N} \Bigg(& \sum_{j=1}^{k} w'_{y_n} \log(p(y_n|\boldsymbol{P'}_j, \boldsymbol{\theta})) \\ & + \sum_{j=k+1}^{m} w'_0 \log(p(y=0|\boldsymbol{P'}_j, \boldsymbol{\theta})) \Bigg) + \frac{\lambda}{2}\|\boldsymbol{\theta}\|^2, \end{aligned} \tag{7}$$

where w'_{y_n} is the empirical estimation of y_n based on patch labels

$$w'_1 = \frac{k \times N_{pos}}{m \times N}, \quad \text{and} \quad w'_0 = 1 - w'_1, \tag{8}$$

where N_{pos} is the number of positive mammograms and N is the total number of mammograms.

One advantage of the label assignment-based multi-instance learning is that it explores all the patches to train the model. Essentially it acts a kind of data augmentation which is an effective technique to train deep networks when the training data is scarce. From the sparsity perspective, the optimization problem of label assignment-based multi-instance learning is exactly a k-sparse problem for the positive data points, where we expect $\{r'_1, r'_2, ..., r'_k\}$ being 1 and $\{r'_{k+1}, r'_{k+2}, ..., r'_m\}$ being 0. The disadvantage of label assignment-based multi-instance learning is that it is hard to estimate the hyper-parameter k. In our experiment, we choose k based on cross validation. Thus, a relaxed assumption for the multi-instance learning or an adaptive way to estimate the hyper-parameter k is preferred.

3.3 Sparse Multi-instance Learning

From the mass distribution, the mass typically comprises about 2% of the whole mammogram on average (Fig. 2), which means the mass region is quite sparse in the whole mammogram. It is straightforward to convert the mass sparsity to the malignant mass sparsity, which implies that $\{r'_1, r'_2, ..., r'_m\}$ is sparse in the whole mammogram classification problem. The sparsity constraint means we expect the malignant probability of part patches r'_i being 0 or close to 0, which is equivalent to the second assumption

in the label assignment-based multi-instance learning. Analogously, we expect r'_1 to be indicative of the true label of mammogram $\boldsymbol{I}$.

After the above discussion, the loss function of the sparse multi-instance learning problem can be defined as

$$\mathcal{L}_{sparse} = \sum_{n=1}^{N} \big(- w_{y_n} \log(p(y_n|\boldsymbol{I}_n, \boldsymbol{\theta})) + \mu \|\boldsymbol{r}'_n\|_1 \big) + \frac{\lambda}{2}\|\boldsymbol{\theta}\|^2, \tag{9}$$

where $p(y_n|\boldsymbol{I}_n, \boldsymbol{\theta})$ can be calculated in Eq. 3, w_{y_n} is the same as that in the max pooling based multi-instance learning, $\boldsymbol{r}_n = (r'_1, r'_2, ..., r'_m)$ for mammogram $\boldsymbol{I}_n$, $\|\cdot\|_1$ denotes the $\mathcal{L}_1$ norm, μ is the sparsity factor, which is a trade-off between the sparsity assumption and the importance of patch $\boldsymbol{Q}'_1$.

From the discussion of label assignment-based multi-instance learning, this learning is a kind of exact k-sparse problem which can be converted to $\mathcal{L}_1$ constrain. One advantage of sparse multi-instance learning over label assignment-based multi-instance learning is that it does not require assign label for each patch which is hard to do for patches where probabilities are not too large or small. The sparse multi-instance learning considers the overall statistical property of $\boldsymbol{r}$.

Another advantage of sparse multi-instance learning is that, it has different weights for general multi-instance assumption (the first part loss) and label distribution within mammogram (the second part loss), which can be considered as a trade-off between max pooling-based multi-instance learning (slack assumption) and label assignment-based multi-instance learning (hard assumption).

3.4 Whole Mammogram Classification using the Learned Model

From the above discussion of the three deep multi-instance variants, we always assume the largest probability r'_1 should be consistent with the malignant label of whole mammogram $\boldsymbol{I}$. In the inference, we can take p'_1 as predicted malignant probability for whole mammogram $\boldsymbol{I}$

$$p(y = 1|\boldsymbol{I}, \boldsymbol{\theta}) = r'_1. \tag{10}$$

4 Experiments

We validate the proposed model on the most frequently used mammographic mass classification dataset, INbreast dataset [16], as the mammograms in other datasets, such as DDSM dataset [5] and mini-MIAS dataset [19], are of low quality. The INbreast dataset contains 410 mammograms of which 94 contains malignant masses. These 94 mammograms with masses are defined as positive mammograms. Five-fold cross validation is used to evaluate model performance. For each testing fold, we use three folds mammograms for training, and one fold for validation to tune the hyper-parameters in the model. The performance is reported as the average of five testing results obtained from the cross-validation.

For preprocessing, we first use Otsu's method to segment the mammogram [18] and remove the background of the mammogram. To prepare the mammograms for following CNNs, we resize the processed mammograms to 224×224. We employ techniques

Table 1. Accuracy Comparisons of the proposed deep multi-instance networks and related methods on test sets.

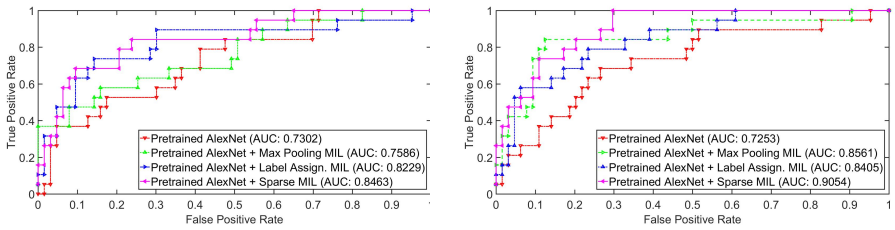
Methodology	Dataset	Set-up	Accu.(%)	AUC(%)
Ball et al. [3]	DDSM	Semi-auto.	87	N/A
Varela et al. [20]	DDSM	Semi-auto.	81	N/A
Domingues et al. [9]	INbr.	Manual	89	N/A
Pretrained CNN [7]	INbr.	Semi-auto.	84±0.04	69±0.10
Pretrained CNN+RF [7]	INbr.	Semi-auto.	$\mathbf{91 \pm 0.02}$	76±0.23
AlexNet	INbr.	Auto.	78.30±0.02	66.80±0.07
Pretrained AlexNet	INbr.	Auto.	80.50±0.03	73.30±0.03
AlexNet+Max Pooling MIL	INbr.	Auto.	83.66±0.02	73.62±0.05
Pretrained AlexNet+Max Pooling MIL	INbr.	Auto.	86.10±0.01	81.51±0.05
AlexNet+Label Assign. MIL	INbr.	Auto.	84.16±0.03	76.90±0.03
Pretrained AlexNet+Label Assign. MIL	INbr.	Auto.	86.35±0.02	82.91±0.01
Pretrained AlexNet+Sparse MIL	INbr.	Auto.	87.11±0.03	83.45±0.05
Pretrained AlexNet+Sparse MIL+Bagging	INbr.	Auto.	90.00±0.02	$\mathbf{85.86 \pm 0.03}$

to augment our data. For each training epoch, we randomly flip the mammograms horizontally, shift within 0.1 proportion of mammograms horizontally and vertically, rotate within 45 degree, and set 50×50 square box as 0. In experiments, the data augmentation is essential for us to train the deep networks.

For the CNN network structure, we use AlexNet and remove the fully connected layers [13]. Through the CNN, the mammogram of size 224×224 becomes 256 6×6 feature maps. Then we use steps in Sec. 3 to do multi-instance learning (MIL). We use Adam optimization with learning rate 0.001 for training from scratch and 5×10^{-5} for training models pretrained on the Imagenet [2]. The λ for max pooling-based and label assignment-based multi-instance learning are 1×10^{-5}. The λ and μ for sparse multi-instance learning are 5×10^{-6} and 1×10^{-5} respectively. For the label assignment-based deep multi-instance network, we select k from $\{4, 8, 12, 16\}$ based on the validation set.

We firstly compare our methods to previous models validated on DDSM dataset and INbreast dataset in Table 1. Previous hand-crafted feature-based methods required manually annotated detection bounding box or segmentation ground truth [3,20,9]. Pretrained CNN used two CNNs to detect the mass region and segment the mass, followed by a third CNN pretrained by hand-crafted features to do the actual mass classification on the detected ROI region [7]. Pretrained CNN+RF further used random forest and obtained 7% improvement. These methods are either manually or semi-automatically, while our methods are totally automated and do not reply on any human designed features or extra annotations.

From Table 1, we observe the models pretrained on Imagenet, Pretrained AlexNet, Pretrained AlexNet+Max Pooling MIL, and Pretrained AlexNet+Label Assign. MIL, improved 2%, 3%, 2% for AlexNet, max pooling-based deep multi-instance learning (AlexNet+Max Pooling MIL) and label assignment-based deep multi-instance learning (AlexNet+Label Assign. MIL) respectively. This shows the features learned on natural images are helpful for the learning of mammogram related deep network. The label assignment-based deep multi-instance networks trained from scratch obtains better performance than the pretrained CNN using 3 different CNNs and detection/segmentation



(a) (b)

Fig. 3. The ROC curve on fold 2 (a) and fold 4 (b) using pretrained AlexNet, pretrained AlexNet with max pooling multi-instance learning, pretrained AlexNet with label assigned multi-instance learning, pretrained AlexNet with sparse multi-instance learning. The proposed deep multi-instance networks improve greatly over the baseline pretrained AlexNet model.

annotation in the training set. This shows the superiority of our end-to-end deep multi-instance networks for whole mammogram classification. According to the accuracy metric, the sparse deep multi-instance network is better than the label assignment-based multi-instance network, and label assignment-based multi-instance network is better than the max pooling-based multi-instance network. This result is consistent with our previous discussion that the label assignment assumption is more efficient than max pooling assumption and sparsity assumption benefited from not having the hard constraints of the label assignment assumption. We obtained different models by using different validation sets for each test fold and used bagging (voting or average different models' predictions) alleviating overfitting to boost the accuracy. Competitive performance to random forest-based pretrained CNN is achieved.

Due to the imbalanced distribution of the dataset where malignant mammograms are only 20% of total mammograms, the receiver operating characteristic (ROC) curve is a better indicator of performance. We compare the ROC curve on test sets fold 2 and fold 4 in Fig. 3 and calculate the averaged area under curve (AUC) of the five test folds in Table 1.

From Fig. 3 and Table 1, we observe that the sparse deep multi-instance network provides the best AUC, and label assignment-based deep multi-instance network obtains the second best AUC. The deep multi-instance network improves greatly over the baseline models, pretrained AlexNet and AlexNet learned from scratch. The pretraining on Imagenet, Pretrained AlexNet, Pretrained AlexNet+Max Pooling MIL, Pretrained AlexNet+Label Assign. MIL, increases performance of AlexNet, max pooling-based deep multi-instance network (AlexNet+Max Pooling MIL), and label assignment-based deep multi-instance network (AlexNet+Label Assign. MIL) by 7%, 8% and 6% respectively. This shows the effectiveness and transferability of deep CNN features learned from natural images to medical images. Our deep networks achieves the best AUC result which proves the superior performance of the deep multi-instance networks.

The main reasons for the superior results using our models are as follows. Firstly, data augmentation is an important technique to increase scarce training datasets and proved useful here. Secondly, our models fully explored all the patches to train our

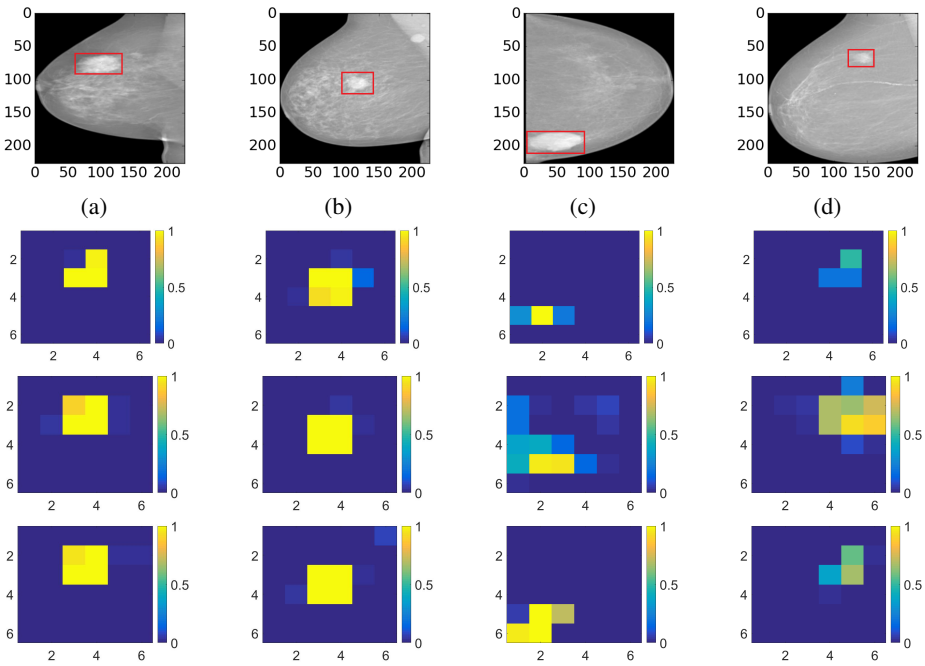


Fig. 4. The visualization of predicted malignant probabilities for instances/patches in four different resized mammograms. The first row is the resized mammogram. The red rectangle boxes are mass regions from the annotations on the dataset. The color images from the second row to the last row are the predicted malignant probability from linear regression layer for (a) to (d) respectively, which are the malignant probabilities of patches/instances. Max pooling-based, label assignment-based, sparse deep multi-instance networks are in the second row, third row, fourth row respectively. Max pooling-based deep multi-instance network misses some malignant patch for mammogram (a), (c) and (d). Label assignment-based deep multi-instance network mis-classifies patches into malignant in (d).

deep networks thereby eliminating any possibility of overlooking malignant patches by only considering a subset of patches. This is a distinct advantage over previous networks that employed several stages consisting of detection and segmentation networks.

5 Discussions

To further understand our deep multi-instance networks, we visualize the responses of linear regression layer for four mammograms on test set, which represents the malignant probability of each patch, in Fig. 4.

From Fig. 4, we can see the deep multi-instance network learns not only the prediction of whole mammogram, but also the prediction of malignant patches within the whole mammogram. Our models are able to learn the mass region of the whole mammogram without any explicit bounding box or segmentation ground truth annotation of

the training data. The max pooling-based deep multi-instance network misses some malignant patches in (a), (c) and (d). The possible reason is that it only considers the patch of max malignant probability in the training and the model is not well learned for all the patches. The label assignment-based deep multi-instance network mis-classifies some patches in (d). The possible reason is that the model sets a constant k for all the mammograms, which causes some misclassification for small mass. One of the potential applications of our work is that these deep multi-instance learning networks could be used to do weak mass annotation automatically, which is important for computer-aided diagnosis.

6 Conclusion

In this paper, we proposed end-to-end trained deep multi-instance networks for whole mammogram classification. Different from previous works using segmentation or detection annotations, we conducted mass classification based on whole mammogram directly. We convert the general multi-instance learning assumption to label assignment problem after ranking. Due to the sparsity of masses, sparse multi-instance learning is used for whole mammogram classification. We explore three schemes of deep multi-instance networks for whole mammogram classification. Experimental results demonstrate more robust performance than previous work even without detection or segmentation annotation in the training.

In future works, it is promising to extend the current work by: 1) incorporating multi-scale modeling such as spatial pyramid to further improve whole mammogram classification, 2) adaptively estimating the parameter k in the label assignment-based multi-instance learning, and 3) employing the deep multi-instance learning to do annotation or provide potential malignant patches to assist diagnoses. Our method should be generally applicable to other bio-image analysis problems where domain expert knowledge and manual labeling required, or region of interest is small and/or sparse relative to the whole image. Our end-to-end deep multi-instance networks are also suited for the large datasets and expected to have improvement if the big dataset is available.

References

1. Andrews, S., Hofmann, T., Tsochantaridis, I.: Multiple instance learning with generalized support vector machines. In: AAAI/IAAI. pp. 943–944 (2002)
2. Ba, J., Kingma, D.: Adam: A method for stochastic optimization. ICLR (2015)
3. Ball, J.E., Bruce, L.M.: Digital mammographic computer aided diagnosis (cad) using adaptive level set segmentation. In: 2007 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society. pp. 4973–4978. IEEE (2007)
4. Beura, S., Majhi, B., Dash, R.: Mammogram classification using two dimensional discrete wavelet transform and gray-level co-occurrence matrix for detection of breast cancer. Neurocomputing 154, 1–14 (2015)
5. Bowyer, K., Kopans, D., Kegelmeyer, W., Moore, R., Sallam, M., Chang, K., Woods, K.: The digital database for screening mammography. In: Third international workshop on digital mammography. vol. 58, p. 27 (1996)

6. Carneiro, G., Nascimento, J., Bradley, A.P.: Unregistered multiview mammogram analysis with pre-trained deep learning models. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 652–660. Springer (2015)
7. Dhungel, N., Carneiro, G., Bradley, A.P.: The automated learning of deep features for breast mass classification from mammograms. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 106–114. Springer (2016)
8. Dietterich, T.G., Lathrop, R.H., Lozano-Pérez, T.: Solving the multiple instance problem with axis-parallel rectangles. Artificial intelligence 89(1), 31–71 (1997)
9. Domingues, I., Sales, E., Cardoso, J., Pereira, W.: Inbreast-database masses characterization. XXIII CBEB (2012)
10. Hou, L., Samaras, D., Kurc, T.M., Gao, Y., Davis, J.E., Saltz, J.H.: Patch-based convolutional neural network for whole slide tissue image classification. arXiv preprint arXiv:1504.07947 (2015)
11. Jiao, Z., Gao, X., Wang, Y., Li, J.: A deep feature based framework for breast masses classification. Neurocomputing 197, 221–231 (2016)
12. Kotzias, D., Denil, M., de Freitas, N., Smyth, P.: From group to individual labels using deep features. In: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 597–606. KDD '15, ACM, New York, NY, USA (2015), http://doi.acm.org/10.1145/2783258.2783380
13. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems. pp. 1097–1105 (2012)
14. Kwok, J.T., Cheung, P.M.: Marginalized multi-instance kernels. In: IJCAI. vol. 7, pp. 901–906 (2007)
15. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proceedings of the IEEE 86(11), 2278–2324 (1998)
16. Moreira, I.C., Amaral, I., Domingues, I., Cardoso, A., Cardoso, M.J., Cardoso, J.S.: Inbreast: toward a full-field digital mammographic database. Academic radiology 19(2), 236–248 (2012)
17. Oeffinger, K.C., Fontham, E.T., Etzioni, R., Herzig, A., Michaelson, J.S., Shih, Y.C.T., Walter, L.C., Church, T.R., Flowers, C.R., LaMonte, S.J., et al.: Breast cancer screening for women at average risk: 2015 guideline update from the american cancer society. Jama 314(15), 1599–1614 (2015)
18. Otsu, N.: A threshold selection method from gray-level histograms. Automatica 11(285-296), 23–27 (1975)
19. Suckling, J., Parker, J., Dance, D., Astley, S., Hutt, I., Boggis, C., Ricketts, I., Stamatakis, E., Cerneaz, N., Kok, S., et al.: The mammographic image analysis society digital mammogram database. In: Exerpta Medica. International Congress Series. vol. 1069, pp. 375–378 (1994)
20. Varela, C., Timp, S., Karssemeijer, N.: Use of border information in the classification of mammographic masses. Physics in Medicine and Biology 51(2), 425 (2006)
21. Wu, J., Yu, Y., Huang, C., Yu, K.: Deep multiple instance learning for image classification and auto-annotation. In: CVPR. pp. 3460–3469. IEEE (2015)
22. Yan, Z., Zhan, Y., Peng, Z., Liao, S., Shinagawa, Y., Zhang, S., Metaxas, D.N., Zhou, X.S.: Multi-instance deep learning: Discover discriminative local anatomies for bodypart recognition. IEEE transactions on medical imaging 35(5), 1332–1343 (2016)
23. Zhou, Z.H., Zhang, M.L., Huang, S.J., Li, Y.F.: Multi-instance multi-label learning. Artificial Intelligence 176(1), 2291–2320 (2012)
24. Zhu, W., Lan, C., Xing, J., Zeng, W., Li, Y., Shen, L., Xie, X.: Co-occurrence feature learning for skeleton based action recognition using regularized deep lstm networks. In: AAAI (2016)
25. Zhu, W., Xie, X.: Adversarial deep structural networks for mammographic mass segmentation. In: preprint (2016)