International Journal of Research in Engineering and Science (IJRES)
ISSN (Online): 2320-9364, ISSN (Print): 2320-9356
www.ijres.org Volume 5 Issue 12 ǁ Dec. 2017 ǁ PP.70-88

Detection from the digitized X-ray mammograms based on the deep active learning

Pragyan Paramita Swain, Anil Kumar Dagar, Flora Das, Kailash Chandra Rout
Department of Electronics and Communication Engineering, NM Institute of Engineering and Technology,Bhubaneswar , Odisha
Department of Electronics and Communication Engineering, Raajdhani Engineering College,Bhubaneswar,Odisha
Department of Electronics and Communication Engineering,Aryan Institute of Engineering and Technology Bhubnaeswar , Odisha
Department of Electronics and Communication Engineering,Capital Engineering College,Bhubaneswar,Odisha

ABSTRACT: Breast mass detection is a challenging task in mammogram, since mass is usually embedded and surrounded by various normal tissues with similar density. Recently, deep learning has achieved impressive performance on this task. However, most deep learning methods require large amounts of well-annotated datasets. Generally, the training datasets is generated through manual annotation by experienced radiologists. However, manual annotation is very time-consuming, tedious and subjective. In this paper, for the purpose of minimizing the annotation efforts, we propose a novel learning framework for mass detection that incorporates deep active learning (DAL) and self-paced learning (SPL) paradigm. The DAL can significantly reduce the annotation efforts by radiologists, while improves the efficiency of model training by obtaining better performance with fewer overall annotated samples. The SPL is able to alleviate the data ambiguity and yield a robust model with generalization capability in various scenarios. In detail, we first employ a few of annotated easy samples to initialize the deep learning model using Focal Loss. In order to find out the most informative samples, we propose an informativeness query algorithm to rank the large amounts of unannotated samples. Next, we propose a self-paced sampling algorithm to select a number of the most informative samples. Finally, the selected most informative samples are manually annotated by experienced radiologists, which are added into the annotated samples for the model updating. This process is looped until there are not enough most informative samples in the unannotated samples. We evaluate the proposed learning framework on 2223 digitized mammograms, which are accompanied with diagnostic reports containing weakly supervised information. The experimental results suggest that our proposed learning framework achieves superior performance over the counterparts. Moreover, our proposed learning framework dramatically reduces the requirement of the annotated samples, i.e., about 20% of all training data.

Keywords:
Breast cancer
Mammography
Mass detection
Deep active learning
Self-paced learning

I. INTRODUCTION

Breast cancer is the most frequently diagnosed cancer and cause of cancer deaths among women worldwide. As American Cancer Society's report in 2017, there were an estimated 252,710 new cases of invasive breast cancer, 63,410 new cases of breast carcinoma in situ, and 40,610 breast cancer deaths among US women [1]. Early and timely detection can lead to a greater range of treatment options to control the development of breast cancer and significantly reduce the mortality. Mammography remains the mainstay of population-based breast cancer screening exam, which can identify more in situ lesions, smaller invasive cancers than other screening methods [2], such as MRI and ultrasound. Mass is the most common type of abnormality in mammogram, and usually appears in relatively dense region. For decades, a range of computer aided diagnosis (CAD) systems designed hand-crafted image features for breast mass detection, by exploiting the properties of shape, size, gradient and texture [3]. However, quite a number of masses are missed and a significant number of false positive tissues are detected, since mass is usually embedded and surrounded by various normal tissues with similar density. Therefore, breast mass detection is still a challenging task.

Recent years, deep learning has achieved impressive break-throughs in various image analysis tasks. For example, Long et al. [4] proposed the fully convolutional networks (FCN) for image segmentation, which token an input image of arbitrary size and produced prediction of spatial density with a same size.

By adapting the structure of FCN, Ronneberger et al. [5] pro-posed U-Net for medical image analysis, which added successive upsampling layers and more skip connections between selected layers. Different from conventional approaches, deep learning approaches automatically learn the optimized features from the raw image, based on the objective functions and the supervised information. As well known, most deep learning approaches require large amounts of well-annotated datasets to optimize the model. The medical image datasets are generated through manual annotation by experienced radiologists. However, there are many difficulties in manual annotation: (1) Manual annotation is very time-consuming, tedious and subjective; (2) We can collect a large mount of data but cannot find resource to annotate them, it is costly to recruit experienced radiologists to annotate large amount of data; (3) The variation of annotation from differ-ent experienced radiologists. Therefore, acquiring well-annotated medical image datasets is the primary challenge.

In practical applications of mass detection, there are large amount of the unannotated samples which usually accompany with weakly supervised information. Active learning (AL) is promising to address the problem that the well-annotated sam-ples are scarce but the unannotated samples are sufficient. Specif-ically, the active learning is an iterative learning method that involves searching for the most informative unannotated samples by query algorithm, selecting them for manual annotation by experienced radiologists, then using the newly annotated samples to update the model at each round. Unlike the conventional supervised learning method, AL employs only a small part of samples that contain the most informative patterns. It can signif-icantly reduce the annotation efforts by experienced radiologists, while improves the efficiency of the model training by obtain-ing better performance with fewer overall annotated samples. Obviously, the query algorithm is the key factor in AL, which is typically performed based on the uncertainty and diversity of samples. The samples with higher uncertainty or unique charac-teristic generally implicate more informativeness for the model updating. Therefore, searching the most informative samples from the unannotated samples is the primary challenge in AL. Besides, the sampling process changes the distribution of the datasets, where the hard samples account for a higher rate in the most informative samples, especially the mass is embedded inside the dense normal tissues in the dense mammograms. Therefore, robustly learning from hard examples (i.e., alleviating the data ambiguity) is another challenging task in AL.

In the course of human learning, humans usually use current experiences to learn new knowledge and rely on the obtained knowledge to accumulate experiences. This interactive process carries out from easy knowledge to complex knowledge gradu-ally, also termed as ‘easy-to-hard’ strategy. Inspired by human being’s learning process, the self-paced learning (SPL) paradigm simulates this process [6,7], in which a model gradually incor-porates easy samples to complex samples into training and thus achieves a more superior model through the constant accumu-lation. It is able to alleviate the data ambiguity and yield a robust model with generalization capability in various scenarios. Intuitively, the criteria of complexity is the key factor in SPL. In this recently rising field, latest studies show the potential of SPL [8–10].

Based on the above motivations, we propose a novel learn-ing framework for breast mass detection that incorporates deep active learning (DAL) and self-paced learning (SPL) paradigm. The DAL denotes the collaboration of deep learning and active learning. Specifically, our proposed learning framework addresses the following challenges: (1) learning more effective feature rep-resentation from only a few of overall annotated samples, and minimizing the annotation efforts by experienced radiologists;

(2) efficiently finding out the most informative samples from the unannotated samples; (3) robustly learning from hard examples (i.e., alleviating the data ambiguity). This work includes the fol-lowing major contributions: First, we develop a novel interactive learning framework for breast mass detection, which provides an efficient learning strategy to obtain better performance with minimum annotation efforts; Second, we develop a novel infor-mativeness query algorithm for DAL, which finds out the most informative samples via considering the uncertainty and diversity simultaneously; Third, we develop a novel self-paced sampling algorithm (i.e., ‘easy-to-hard’ strategy) for SPL, which selects a number of the most informative samples on the bias of com-plexity and pace. In detail, we first employ a few of annotated easy samples to initialize the deep learning model (i.e., FCN) using Focal Loss. In order to find out the most informative samples, we employ the proposed informativeness query algorithm to rank the large amounts of unannotated samples. Next, we employ the proposed self-paced sampling algorithm to select a number of the most informative samples. Finally, the selected most informa-tive samples are manually annotated by experienced radiologists, which are added into the annotated samples for the model up-dating. This process is looped until there are not enough most informative samples in the unannotated samples.

II. RELATED WORK

In this section, we briefly review the latest developments on breast mass detection, active learning and self-paced learning respectively.

Breast mass detection. The conventional mass detection methods depend on the combination of hand-craft features and specific classifiers. Oliver et al. [3] provided a quantitative com-parison for various conventional mass detection methods, and analyzed the advantages and disadvantages of the used strategies qualitatively. The latest developments tend to employ deep learn-ing techniques for detecting, segmenting and classifying breast masses from mammogram. Generally, the deep learning methods achieve superior performance over the conventional methods. For example, Arevalo et al. [11], Kooi et al. [12] and Dhungel et al. [13] proposed staged optimizing strategies for mass classification in their studies, respectively. In the features representation stage, they employ deep learning techniques to automatically extract discriminative features and investigate extent hand-crafted fea-tures complementally. In the classification stage, all the features are fed into classifiers to make final decision. However, the staged optimizing strategies are inefficient. By making use of the state-of-the-art object detection method, Ribli et al. [14] proposed a CAD system for mass detection based on Faster R-CNN [15]. This deep learning framework is appropriate for simultaneously de-tecting, localizing and classifying large objects in high definition and contrast natural images. Al-masni and Al-antari et al. [16,17] proposed a CAD system for mass detection based on You Only Look Once (YOLO), a ROI-based Convolutional Neural Network (CNN). The YOLO-based CAD system can handle detection and classification simultaneously in one framework. However, there are still several limitations of directly applying Faster R-CNN or YOLO for mass detection in mammogram. The object detection methods locate the key points on the small-scale feature maps, and then use multiple scale boxes to generate region proposals on the original image. The region proposals are used for classification in subsequence, thus the boxes cannot be too small. Therefore, the object detection methods cannot accurately detect the small lesions. Recently, many FCN variations has been proposed for lesion detection tasks via incorporating with lesion segmentation task in medical images [18–23]. For example, the U-Net, i.e., a variation of FCN, employs skip connections in its architecture to acquire detailed information from low-level large-scale feature maps, which can improve the recognition capability of small lesions. These studies suggest that FCNs have potential to achieve promising performance in mass detection.

Active learning. The active learning (AL) paradigm focuses on actively selecting the most informative samples, in order to learn better feature representation from only a few of overall annotated samples and to minimize the annotation efforts. The key factor of AL is the sample selection criteria, which usually relies on the measurement of uncertainty and diversity. The uncertainty aims to reduce the expected error of model [24], thus the large amounts of unannotated samples are ignored due to their rela-tively far away from the decision boundary. The diversity [25– 27] aims to enrich the feature representation of model and to enhance the generally discriminative capability, thus the repre-sentatively diverse samples are worthy to be selected. Recently, many latest studies [28–31] employ AL in various scenarios of image analysis, and investigate more effective sample selection criteria. In medical imaging community, there come up a lot of impressive studies [32–36] that employ AL to tackle the problem of scarce well-annotated medical samples. For example, Melendez et al. [32] proposed a lesion detection framework by embedding a multiple instance learning (MIL) classifier within AL. To minimize the annotation efforts, meaningful lesion regions are selected with the help of AL. Because many latest developments employ deep learning techniques to collaborate with AL [34,35,37,38], these frameworks are also termed deep active learning (DAL).

Self-paced learning. At the beginning, Bengio et al. [6] in-troduced the concept of curriculum learning, in which a model mimics human being's learning process by gradually accumulat-ing knowledge from easy to complex. To make it more imple-mentable, Kumar et al. [7] formulated this learning philosophy as an explicit paradigm named self-paced learning (SPL). The SPL paradigm is able to alleviate the data ambiguity and guide a robust learning manner in complex scenarios. For example, Sangineto et al. [8] employed the SPL to select the highest-confidence bounding boxes as pseudo-ground truth in a weakly-supervised scenario of object detection. Specifically, this training strategy discarded noisy training bounding boxes and progres-sively trained a Fast R-CNN [39] using the most likely bound-ing boxes. Zhang et al. [9] also employed SPL to collaborate with multiple instance learning (MIL) in the co-saliency detection framework. Besides, Lin et al. [10] developed a cost-effective deep learning framework for face identification, by combining the active learning (AL) and the self-paced learning (SPL). The framework investigates the resolution of automatically annotat-ing new instances and incorporating with SPL under the weak expert recertification.

III. METHODOLOGY

In this section, we first illustrate our learning framework and explain the mechanism. Next, we introduce the deep learning model in the learning framework. Finally, we specifically describe the proposed informativeness query algorithm and self-paced sampling algorithm, which are the key factors in the learning framework. The workflow of our learning framework is demon-strated in Fig. 1.

3.1. Mechanism of learning framework

The learning framework includes the following steps: Training Initial model, Predicting unannotated samples, Selecting informa-tive samples, Manual annotation, Updating model. Except the first step of Training Initial model, the learning framework iteratively alternates among the other steps. This iterative learning pro-cess is looped until there are no more informative samples in unannotated samples, e.g., less than 2% of the annotated samples.

For convenient presentation, we define the following notation. A_t denotes the annotated samples and U_t denotes the unanno-tated samples at the round t. Specifically, A_0 and U_0 denote initial annotated samples and initial unannotated samples, respectively. D denotes the overall training dataset in our learning framework, where $D = U_t \cup A_t$ and $U_t \cap A_t = \emptyset$. I_t denotes the set of most informative samples which are selected at the round t. Obviously, the annotated samples A_t increase gradually and the unannotated samples U_t decrease. For the purpose of minimizing the annotation efforts, the final annotated samples A_c should much less than the overall training dataset D.

Training Initial model: At the beginning, a few of easy samples are annotated by experienced radiologists for initializing the deep learning model F_0 (i.e., FCN). Specifically, we randomly select a few of samples that have significant visual characteristics of mass and convinced abnormal mass descriptions in their diag-nostic reports. Obviously, these samples are easy to be annotated, i.e., initial annotated samples A_0. Additionally, we employ Focal Loss [40] in the deep learning model to further improve the performance.

Predicting unannotated samples: At the round t, we apply the current deep learning model to predict the unannotated samples U_t , and then generate a set of heatmaps. The large amounts of unannotated samples U_t accompany with diagnostic reports, in which non-uniform analysis and diagnosis (i.e., weakly supervised information) are given by different clinicians.

Selecting informative samples: We rank the unannotated sam-ples U_t using the proposed informativeness query algorithm, which estimates the informativeness of samples via considering the uncertainty and diversity of samples. The uncertainty is measured from the predicted heatmaps and weakly supervised information. The diversity accounts for the uniqueness and rep-resentativeness property among samples. Then, we can find out the most informative samples based on the ranking. We employ the self-paced sampling algorithm to select a number of the most informative samples I_t . The self-paced sampling algorithm tends to preferentially select the 'easy' samples, by considering the complexity of samples.

Manual annotation: The selected most informative samples I_t are annotated manually by experienced radiologists. Specifically, with the help of the diagnostic reports, a group of experienced radiologists manually draw the contours of masses on the mam-mograms and assign the convinced confidence. Then, we can convert the contours into annotated masks. After the manual annotation, the annotated samples are extended, and the newly annotated samples are removed from the unannotated samples.

Updating model: The extended annotated samples are used to update the deep learning model, and then the updated deep learning model F_t is used in the next round.

3.2. Deep learning model

In our proposed learning framework, we employ an end-to-end and pixels-to-pixels deep learning model, named fully convo-lutional networks (FCN) [4]. Formally, FCN transfers classification tasks to segmentation tasks by reinterpreting classification net-works as fully convolutional and fine-tuning from their learned representations. We employ the VGG-16 [41] as the backbone of FCN. Besides, FCN applies transpose convolution, upsampling and skip connection of feature maps in the decoder to make the output regains the same size as input. For its specific architec-ture, the FCN achieves a nonlinear and local-to-global feature representation that embeds the low-level visual features and the

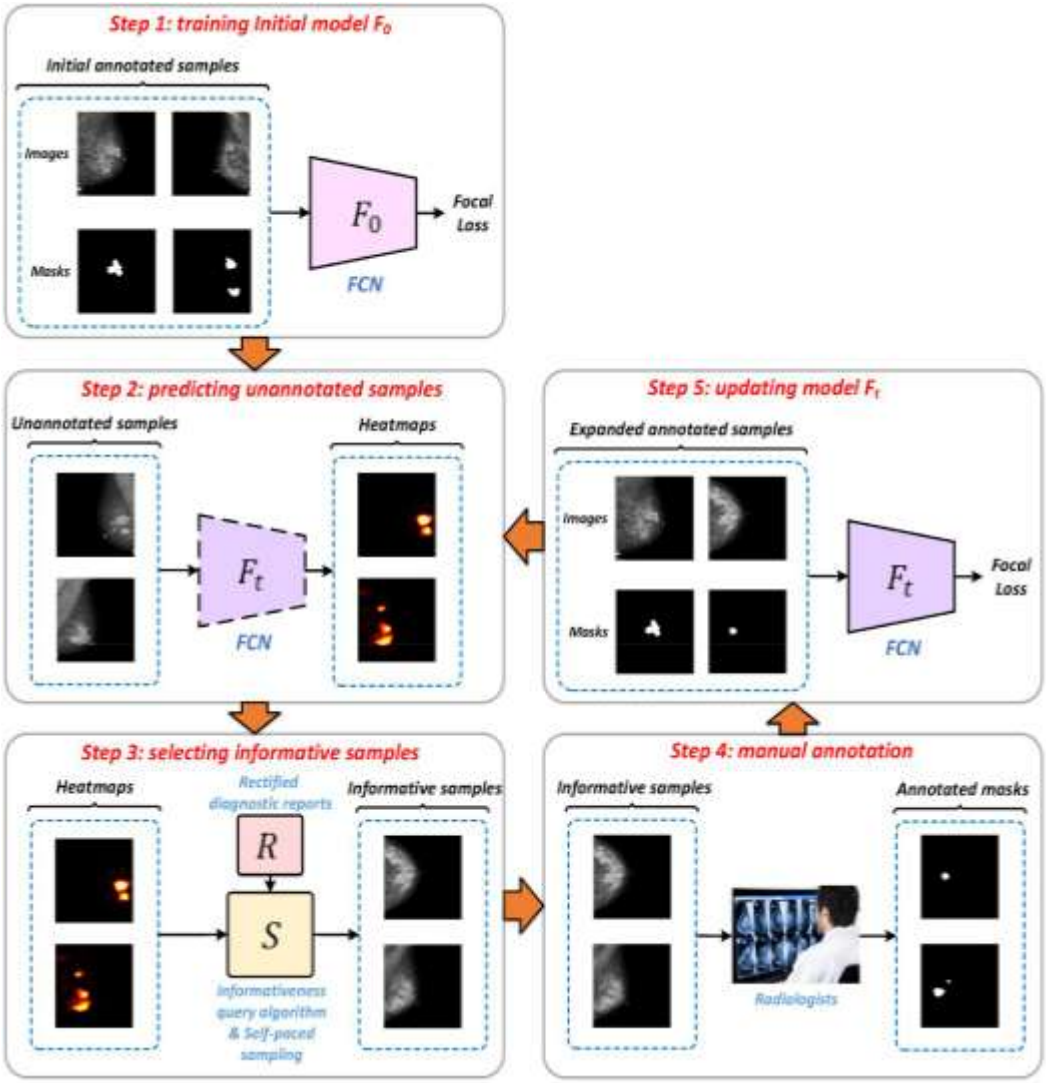


Fig. 1. The workflow of our proposed learning framework. **Step 1. Training Initial model:** a few of annotated easy samples are employed to initialize the deep

learning model using Focal Loss. **Step 2. Predicting unannotated samples:** applying the current deep learning model to the unannotated samples would generate a set of heatmaps. **Step 3. Selecting informative samples:** we employ informativeness query algorithm and self-paced sampling algorithm to select a number of the most informative unannotated samples. **Step 4. Manual annotation:** the selected unannotated samples are annotated manually by experienced radiologists, thus the annotated samples are extended. **Step 5. Updating model:** the extended annotated samples are used to update the deep learning model, and then the updated deep learning model is used in the next round.

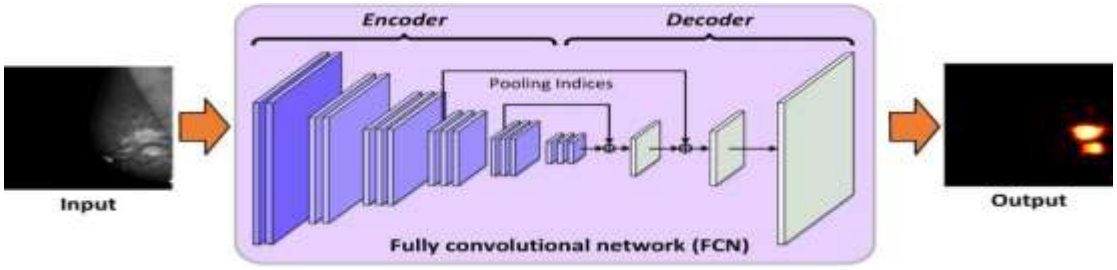


Fig. 2. The architecture of FCN in our learning framework. The output of FCN is heatmap.

high-level semantic features. As shown in Fig. 2, the output of FCN is heatmap, in which the value of each pixel indicates the probability that the corresponding pixel belongs to mass lesion.

Assuming a grayscale image I with pixels $\mathbf{p} = (p_x, p_y)$ and intensities $I(\mathbf{p}) \in V$ on a discrete grid $X \times Y$, where V is the discrete domain of intensity, such as [0, 255]. Additionally, the annotated mask M of this image is available, its pixels correspond to semantic binary class labels $M(\mathbf{p}) \in \xi$, where $\xi = \{0, 1\}$. The training of FCN is to learn the parameters $\{\mathbf{w}, \mathbf{b}\}$ of mapping function F. The heatmap $T(\mathbf{p})$ can be calculated by

$$T(\mathbf{p}) = F(I(\mathbf{p}); \mathbf{w}, \mathbf{b}).$$

Therefore, the loss function $L(\mathbf{w}, \mathbf{b})$ of image I can be written as

$$L(\mathbf{w}, \mathbf{b}) = \frac{1}{N} \sum_{\mathbf{p}} H(M(\mathbf{p}), T(\mathbf{p})),$$

where $T(\mathbf{p}) \in [0, 1]$, N is the number of pixels in image I, the $H(\cdot)$ is the loss function per pixel. The mammogram usually has a high resolution, e.g., 2000 × 3000. It is computationally expensive to entirely feed it into FCN. The promising solution is to resize the mammogram and its mask into a lower resolution, e.g., 1600 × 1600.

In the mass detection task, the class imbalance between foreground and background is extreme that the mass lesions in foreground occupy much less pixels than the normal tissues in background. The class imbalance problem causes machine learning bias toward the majority class, and tends to produce suboptimal models that misclassify the minority class. To address this problem, we employ Focal Loss [40] to train the FCN. Formally, focal loss transforms the standard cross entropy loss into a weighted form, which reduces the relative loss for well-classified pixels and puts more focus on ill-classified pixels. This modification provides a more effective alternate to previous cross entropy loss for dealing with class imbalance.

batch size is set as 4. The training procedure contains 80 epochs in total. Additionally, we employ data augmentation strategy in the FCN training, such as stretch, cropping, flipping and discoloration. Specifically, the data reader loads a batch of samples from the annotated samples, applies the data augmentation strategy, and then inputs the samples into the FCN. The data augmentation strategy can introduce variation of samples to restrain the overfitting.

3.3. Informativeness query algorithm and self-paced sampling algorithm

We now describe how we actively select the most informative samples for manual annotation.

3.3.1. Informativeness query algorithm

We propose the informativeness query algorithm to estimate the informativeness of unannotated samples via considering the uncertainty and diversity, and then to rank the unannotated samples.

The uncertainty is measured from the predicted heatmaps and weakly supervised information. For the samples with higher uncertainty, the current model is not confident to have localized the mass lesions correctly. Therefore, annotating these samples and adding them into annotated samples could reduce the expected error of the updated model. Because the diagnostic reports (i.e.,

Assuming CE(·) is cross entropy loss, where CE(μ, μ˜) is a measure of dissimilarity between μ and μ˜. In binary classification situation, it can be calculated by

CE(μ, μ) μjlog(μj)

˜ = − Σj ˜

= −μlogμ˜ − (1 − μ)log(1 − μ˜),

μ , μ ,

where ∈ {0 1} specifies the truth label and ˜∈[0 the estimated probability of the class with label μ. For notational convenience, we define μc

μc μ,˜μ=1

= { 1−μ,˜μ=0.

Then, we can rewrite the cross entropy loss as

CE(μc log(μc)).

= −

from cross entropy loss, by

The focal loss FL(·) is transformed and multiplying a modulating factor (1 − μc)) . The formal focal loss FL(μc) is defined as

FL(μc) = −αc (1 − μc)γ log(μc),

where the tunable parameter γ ≥ 0, and αc is defined as

αc = { α, μ 1

weighte d factor

weakly supervised information) are usually given by different inclinicians, most of their descriptions are subjective, incomplete, non-uniform, imprecise and low-credibility. In order to effectively use the weakly supervised information, experienced radiologists rectify the diagnostic reports via locating suspicious mass regions ((e.g., bounding boxes) and assigning resilient) confidence.

∈ Ut at the round t, the

For an unannotated sample Ii FCN outputs its heatmap Ti (**p**)

(**p**) = F t(Ii(**p**); **w**, **b**). (9)

Ti

The uncertainty UCi t is measured by

1

UCi t Ti t (**p**) Si(**p**) , (10)

= N Σ**p** −

where Si(**p**) ∈ [0, 1] denotes the weakly supervised information.

i(**p**), the weakly Different from the annotated mask supervised M information provides resilient confidence for suspicious mass regions, e.g., suspicious tuberosity, high aprobability dense region, obscure mass. For convenience, the normal regions are assigned the resilient confidence of 0 and the suspicious mass regions are assigned resilient confidence of [0.5, 1.0]. Besides, the suspicious mass regions should have more weighted contributions for uncertainty than the normal regions. Thus, the uncertainty UCi t is rewritten by

$1-\alpha, \mu^{=} = 0,$

where the tunable parameter $\alpha \in (0, 1)$ and $\gamma \geq$ control the weights between classes, e.g., $\gamma = 2, \alpha = 0.75$.

Thus, the loss function of FCN for image I can be written as

$$L(\mathbf{w}, \mathbf{b}) = \frac{1}{N} \sum_{\mathbf{p}} FL(\mu_c^{(\mathbf{p})}) = \frac{1}{N} \sum_{\mathbf{p}} \alpha_c (1 - \mu^{(\mathbf{p})})^{\gamma} \log(\mu^{(\mathbf{p})}).$$

To train the FCN, we employ Adam [42] optimizer to perform back-propagation, where the learning rate is fixed to 10^{-5} and the

$t = \lambda^{+} t(\mathbf{p}^{+}) - S_i(\mathbf{p}^{+})$

$\sum \lambda^{-} ( ) ((11)$

$+ t \mathbf{p} - S_i ^{-}),$

$N \mathbf{p}^{-} {}^{T}i \mathbf{p}$

where $\lambda^{+} \gg \lambda^{-}$, $\mathbf{p}^{+}$ and $\mathbf{p}^{-}$ denote pixels in suspicious mass regions and normal regions, respectively.

Note that, due to the unreliability of the weakly supervised information, the uncertainty provides an approximate quantitative evaluation to find more uncertain samples. For example, two suspicious mass regions have the same predicted value of 0.7, and different resilient confidence of 0.8 and 0.9 in the weakly supervised information. Obviously, the suspicious mass region with resilient confidence of 0.9 has higher uncertainty.

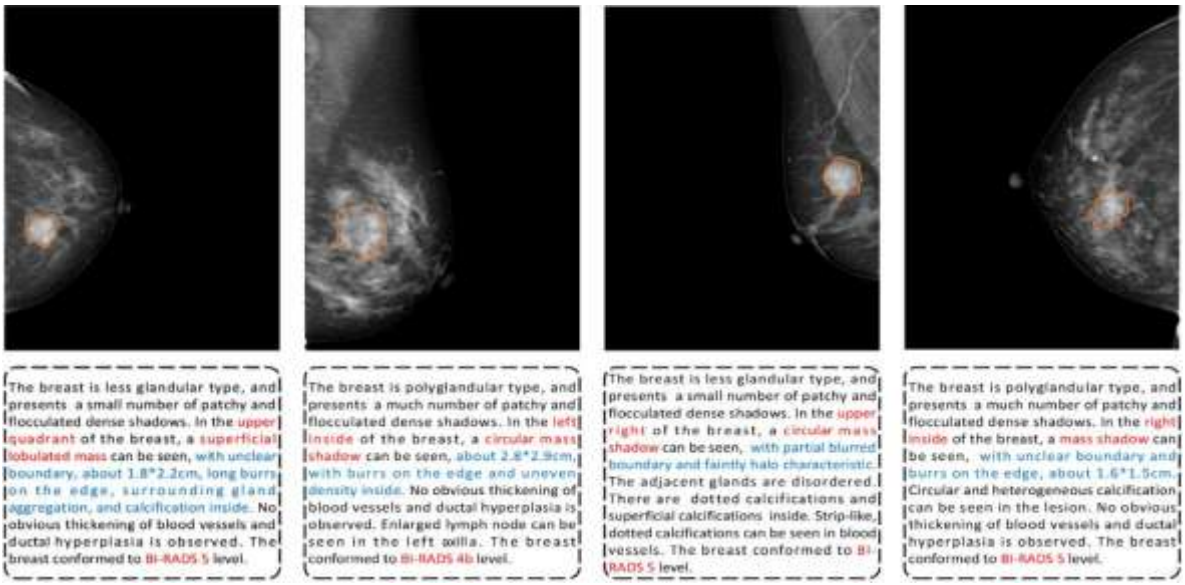


Fig. 3. There are four annotated samples and their diagnostic reports. The orange contours are manually annotated by experienced radiologists. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

The diversity accounts for the uniqueness and representativeness property among samples. The samples with high diversity are more informative that enrich the feature space of samples and enhance the generally discriminative capability of the model.

For an unannotated sample $I_i \in U_t$ at the round t, we can calculate its diversity DS_i by

$$DS_i = 1 - \frac{1}{N_u^t} \sum_{I_j \in U_t \setminus I_i} \psi(I_i, I_j), \quad (12)$$

where N_u^t denotes the number of unannotated samples U_t, $\psi(\cdot)$ is a similarity function. Therefore, similar samples do not provide more information for updating the model. The purpose of diversity is to select samples with high diversity, and not to select similar samples.

The proposed informativeness query algorithm considers both the uncertainty and the diversity where $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$ denotes the training data, $L(w; \mathbf{x}_i, y_i)$ denotes the loss function between the label y_i and the estimated one, ζ) denotes a self-paced regularizer, $v = [v_1, v_2, \ldots, v_n]$ denote the weight variables of the samples, ζ is a parameter of the learning pace (i.e., round t,round).

Interestingly, the SPL has the opposite sampling criteria as the active learning. The SPL tends to select 'easy' samples with lower differences $L(w; \mathbf{x}_i, y_i)$, where $\lim_{L_i \to 0} v_i = 1$. Based on the uncertainty, the AL tends to select samples with relatively large differences. This finding inspires us to investigate the compatibility between the two learning paradigms and the possibility of making them complementary to each other.

Furthermore, as the Eq. (14), the labels y_i (i.e., annotated masks) of the unannotated samples are not available. As previously mentioned, the weakly supervised information $S_i(\mathbf{p})$ is not convinced, thus using $S_i(\mathbf{p})$ as y_i is biased to calculate $L(w; \mathbf{x}_i, y_i)$.

In this situation, we introduce the complexity c_i as the

simultaneously. We further define the informativeness QA_i^t that combines the uncertainty UC_i^t and the diversity DS_i^t. The informativeness QA_i^t measures how informative the unannotated sample can provide to current model. The informativeness QA_i^t can be calculated by

$$QA_i^t = \eta * UC_i^t + (1-\eta) * DS_i^t, \quad (13)$$

where $\eta \in (0, 1)$ is a coefficient to weight the uncertainty UC_i^t and the diversity DS_i^t. Specifically, the more unannotated samples U_t should be assigned smaller η to emphasize the generally discriminative capability. According to the informativeness QA_i^t, the unannotated samples can be ranked and the most informative top k_t samples are obtained.

3.3.2. Self-paced sampling algorithm

Kumar et al. [7] formulated the philosophy of self-paced learning (SPL), which employed a 'easy-to-hard' strategy to facilitate learning. This strategy preferentially selects 'easy' samples that have high confidence. Formally, Jiang [43] provided more comprehensive understanding of SPL as a general optimization problem

$$\min_{w,\, v\in[0,1]^n} \sum_{i=1}^{n} v_i L(w; x_i, y_i) + r(v; \zeta), \quad (14)$$

sampling measurement. We propose the self-paced sampling algorithm to select a number of the most informative samples. Obviously, the complexity is strongly correlated with the weakly supervised information. The more convinced supervised information signifies the lower complexity, thus the possibility of being selected v_i increases with the lower complexity $\lim_{c_i\to 0} v_i = 1$. As the aforementioned weakly supervised information $S_i(\mathbf{p})$, we define the complexity $c_i \in [0, 1]$ as

$$1 - c_i \propto \frac{1}{\overline{N}_{\mathbf{p}}} \sum S_i(\mathbf{p}) = \frac{\tau_i}{\overline{N}_{\mathbf{p}^+}} \sum S_i(\mathbf{p}^+), \quad (15)$$

where τ_i denotes a normalization factor. This definition indicates that more convinced weakly supervised information corresponds to more 'easy' samples. For convenience, we take a binary of $v_i \in \{0, 1\}$ by a probability sampling function $P(\cdot)$. Then, the v_i can be written by

$$v_i = P(1 - c_i, \zeta). \quad (16)$$

As definition of ζ in Eq. (14), the learning pace parameter ζ in Eq. (16) states that we should incorporate more hard samples for training when the learning pace gets larger. Therefore,

$$\lim v_i = 1,\ i \in \{1, \dots, n\}. \quad (17)$$

In brief, we first employ the informativeness query algorithm to find out the most informative samples, then employ the self-paced sampling algorithm to select a number of relatively easy samples from the most informative samples. When the round t increases, the most informative samples will be less and less, while the self-paced sampling algorithm will select higher pro-portion of the most informative samples.

IV. EXPERIMENTS

4.1. Datasets

We employ a database with 2,223 standard full-field digital mammograms (FFDM), which are accompanied with diagnostic reports. These mammograms are chose on the basis of suspicious mass

descriptions in diagnostic reports. All the mammograms come from four standard views: a cranial-caudal (CC) view and a mediolateral-oblique (MLO) view, from both the left and right breasts. We collect the database from the First People's Hos-pital of Xiang Yang, Affiliated Hospital of Hubei University of Medicine. These mammograms are acquired using 2 kind of FFDM devices from different manufacturers (i.e., Planmed Nuance and IMS Giotto) with spatial resolution of 85 μm/pixel.

We randomly split the database into a training dataset with 1,912 mammograms and a test dataset with 311 mammograms. The training–test ratio of our database is approximately equal to the training–test ratio of the DDSM [44] dataset. Based on the philosophy of our learning framework, only a small part of samples in the training dataset are selected for manual annota-tion. Specifically, the initial annotated samples A_0 contain 220 'easy' mammograms, thus the initial unannotated samples U_0 contain 1,692 mammograms. Four annotated samples and their diagnostic reports are demonstrated in Fig. 3. Besides, the process of the manual annotation by experienced radiologists conforms to the standardized specifications of the public INBreast [45] dataset. In order to guarantee the consistent, the annotations are consulted by 5 experienced radiologists.

As the related works for mass detection [3,12,13], the most appropriate metric for lesion detection evaluation is FROC [46] curve. FROC curve is defined as the plot of true positive rate (i.e., TPR, recall) versus the average number of false positives per image (FPI) at multiple thresholds, where the IoU is set as 0.2. The true positive rate (TPR) represents the accuracy of mass, and the average number of false positives per image (FPI) represents the average number of misclassified normal tissues in each image. We further employ the true positive rate (TPR) at given false positives per image (FPI) as another metric, such as TPR@2.0FPI. In addition, we calculate the partial area under the FROC curve (PAUC) to further investigate the difference quantitatively. For convenient comparison, we choose a fixed FPI range of [0.5, 2.0] to calculate the PAUC in the FROC curve. Because the TPR is close to saturation when the FPI is greater than 2.0, and the threshold is close to 1 when the FPI is less than 0.5.

The experiments are implemented on a workstation with 6 CPU cores and 2 NVIDIA TESLA P40 GPUs, each GPU has 24GB memory. We employ Keras [47] with the TensorFlow [48] backend as our deep learning framework.

4.2. Comparison between focal loss and cross entropy

Experiments have been conducted to compare the focal loss (FL) and cross entropy (CE). We employ the initial annotated samples to train two models, which have the same network architecture but different loss functions, i.e., focal loss (FL) and cross entropy (CE). Then, we evaluate these models on the test dataset. Fig. 4 shows the heatmaps and overlays of a test samples.

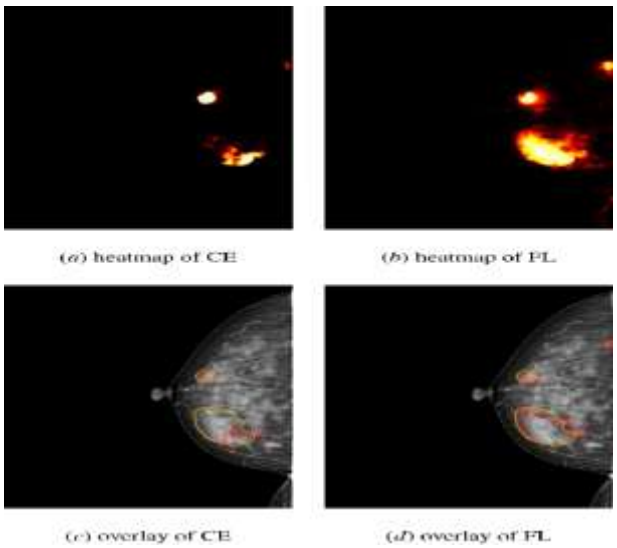


Fig. 4. The heatmaps and overlays of a test samples. In subfigure c and d, the orange counters denote the manual annotations and the red counters denote the predicted suspicious mass lesions using the default threshold of 0.5. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

As shown in Fig. 4, the heatmap of FL pays more attention to the more hard regions, where the confidence in the heatmap aggregates in the moderate threshold range. In the contrary, the heatmap of CE tends to predict high confidence (i.e., 0 or

1) for these regions, but could lead to more misclassification. The suspicious mass lesions can be detected by taking different thresholds for binarization, Fig. 4(c) and Fig. 4(d) demonstrate the suspicious mass lesions (red color) predicted by two models using the default threshold of 0.5, respectively. For this sample, two models all recall the suspicious mass lesions, but the under suspicious mass lesion predicted by the model with CE is much less than the manual annotation. In other words, the model with FL provides more robust performance than the model with CE.

The FROC curves of this comparison are shown in Fig. 5. Obviously, we can observe that the model with FL outperforms the model with CE. At the same FPI range of [0.5, 2.0], the two models obtain 0.8515 PAUC, 0.8780TPR@2.0FPI and 0.8398 PAUC, 0.8780TPR@2.0FPI, respectively. Although the model with FL presents better PAUC performance, it still remains a weakness that the predicted suspicious mass lesions could be much larger than manual annotations in lower thresholds. As the FROC ten-dency of the model with FL in Fig. 5, the TPR tends to saturation and even suffers a little degradation in lower thresholds.

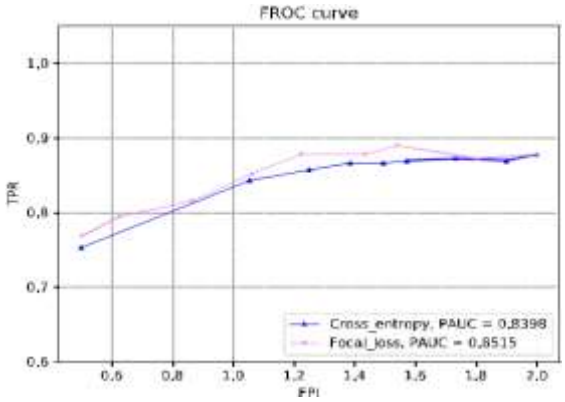


Fig. 5. Performance comparison between the models with focal loss (FL) and cross entropy (CE).

4.3. Evaluations of informativeness query algorithm

In practice, the informative samples from the unannotated samples can be grouped into 4 clusters qualitatively by informa-tiveness query algorithm, e.g., 'mismatching', 'none-recall', 'lower-recall', 'over-FP'. In the 'mismatching' cluster, the mass lesions are not recalled and a number of false positive regions are recalled. In the 'none-recall' cluster, neither mass lesions nor false positive regions are recalled. In the 'lower-recall' cluster, the mass lesions are located but the recalled regions are only a small part of mass lesions. In the 'over-FP' cluster, the mass lesions and a number of false positive regions are recalled. According to the Eq. (11), the samples in 'mismatching' and 'none-recall' should have higher uncertainty. In generally, the samples in 'none-recall' and 'lower-recall' are relatively less, thus should have higher di-versity based on the Eq. (12). Therefore, according to the Eq. (13), the samples in 'mismatching' and 'none-recall' are usually the most informative samples.

We conduct a set of experiments to verify our proposed infor-mativeness query algorithm. We employ the FCN with focal loss trained on the initial annotated samples as the initial model, then generate a set of heatmaps of the initial unannotated samples. Certainly, the most of initial unannotated samples (i.e., 1,140 in 1,692) are classified as uninformative samples, which mass lesions are correctly detected with few false positive regions. According to the Eq. (13), the ranks of these samples are low. Therefore, we focus on the high rank samples (i.e., 552 in 1,692), which can be split into 4 clusters, i.e., 154 samples of 'mismatch', 47 samples of 'none-recall', 62 samples of 'lower-recall' and 289 samples of 'over-FP'. Four typical samples from different clusters with their heatmaps are illustrated in Fig. 6.

In order to verify the improvement of model updating by these clusters independently, we select 98 samples from 'mismatching', 39 samples from 'none-recall', 55 samples from 'lower-recall', 110 samples from 'over-FP', respectively. This selection considers both the uncertainty and the diversity simultaneously, and ignores a number of hard samples (i.e., indecipherable samples). These selected samples are manually annotated by a group of experi-enced radiologists. Then, we obtain 4 sets, i.e., Active_c1, Active_c2, Active_c3 and Active_c4, by adding the newly annotated samples into the initial annotated samples, respectively. We employ these 4 sets to update the initial model, respectively. Then, we evaluate the updated models on the test dataset. The FROC curves of de-tection results are shown in Fig. 7. Obviously, the updated models on Active_c1 and Active_c2 both obtain significant improvement

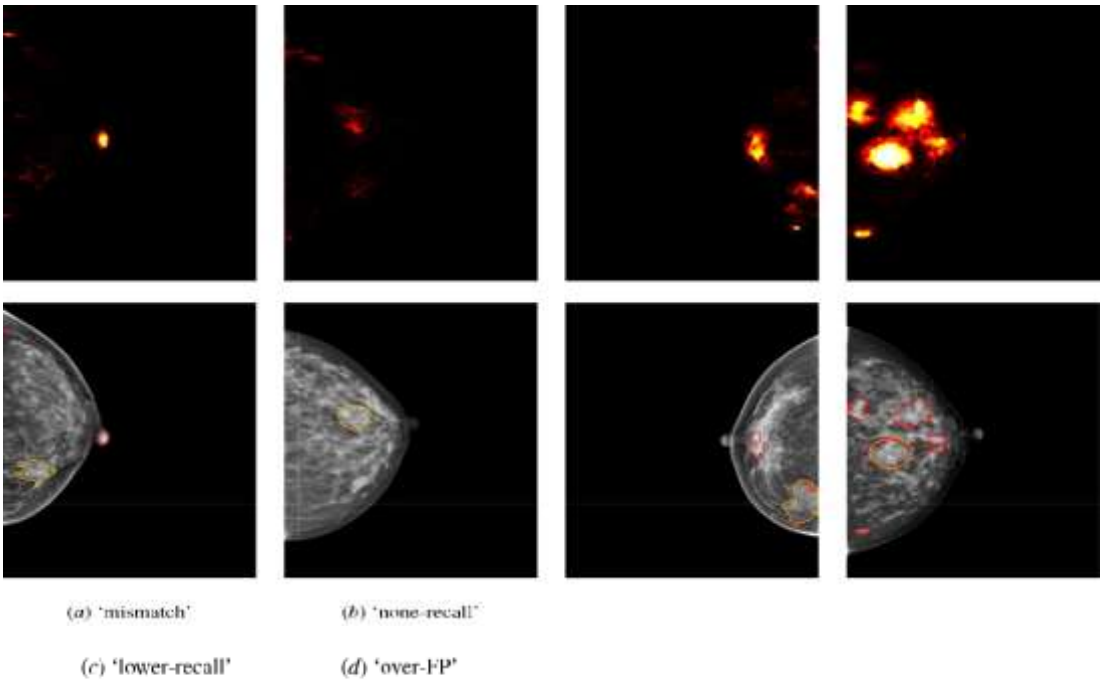


Fig. 6. Four typical samples from different cluster, i.e., 'mismatch', 'none-recall', 'lower-recall' and 'over-FP'. The upper row shows the heatmaps of these samples. The lower row shows the overlays of predicted suspicious mass regions and manual annotations on original images. Note that, the orange contours denote the manual annotations and the red contours denote the predicted suspicious mass lesions. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

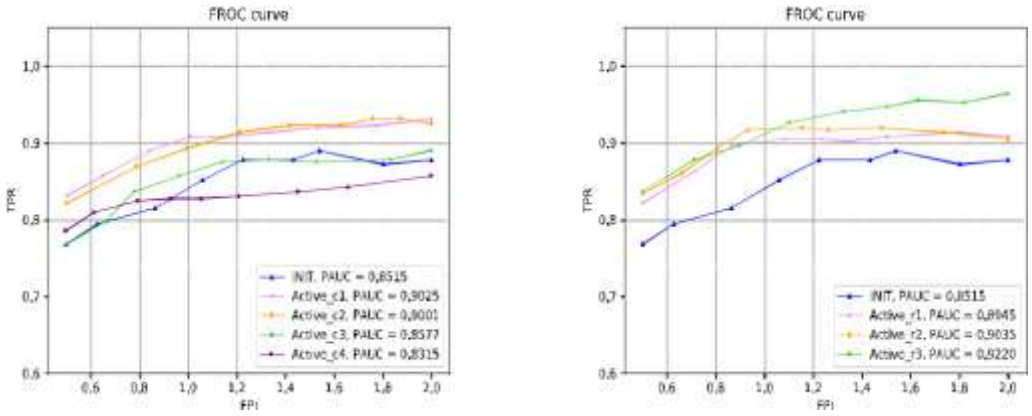


Fig. 7. Performance comparison between various clusters of informative samples.

than the initial model (i.e., 0.8515 PAUC), where 0.9025 PAUC and 0.9001 PAUC are achieved at the same FPI range of [0.5, 2.0], respectively. Therefore, the samples in 'mismatching' and 'none-recall' are the most informative samples. Besides, the updated model on Active_c3 obtains very few improvement than the initial model. Interestingly, the updated model on Active_c4 suffers a lit-tle degradation than the initial model. Because the added samples from 'over-FP' contain several times false positive regions than the suspicious mass regions, which aggravate the class imbalance problem. In addition, the false positive regions have high similar-ity with the suspicious mass regions, and increase the difficulty of learning.

4.4. Evaluations of self-paced sampling algorithm

As the demonstration in previous experiments, the samples in 'mismatching' and 'none-recall' are the most informative samples. In our learning framework, a number of these most informa-tive samples are selected by self-paced sampling. The selected most informative samples are manually annotated and added into the annotated samples. The current model is updated using the extended annotated samples in each round. Fig. 8

shows the FROC curves of detection results on the test dataset, where Active_r1, Active_r2 and Active_r3 denote the detection results of three rounds. In the three rounds, we select 137, 27, 18 most in-formative samples, respectively. Therefore, 182 most informative samples are selected in total, and 402 samples are annotated in total (about 20% of all training data). Our iterative learning frame-work is terminated in the third round, because of not enough most informative samples in next round (i.e., less than 2% of the current annotated samples). We can observe that a consistent improvement is achieved in each round. At the same FPI range of [0.5, 2.0], the updated model on Active_r3 presents the best 0.9220 PAUC and 0.9643 TPR@2.0FPI, as shown in Table 1. An-other interesting detail about Fig. 8 is that the updated models on Active_r1, Active_r2 and Active_r3 obtain very close performance in the low FPI range (e.g., < 1.2) and relatively large differences in the high FPI range. This finding indicates that many illegible mass lesions are recalled in lower thresholds and the improvement of subsequent rounds comes from hard samples which masses are embedded inside the dense tissues.

4.5. Comparison with the counterparts

4.5.1. Active learning (AL)

We employ active learning as the first counterpart, which is the state-of-the-art methods for minimizing the annotation

Fig. 8. Performance comparison between the initial model and the updated models in our learning framework.

Table 1 Experimental results of our proposed method and counterparts on test dataset.

Experiments	Name	PAUC	TPR@2.0FPI	Annotated samples
FL vs. CE	CE	0.8398	0.8780	220
	FL(Baseline)	0.8515	0.8780	220
Evaluations of IQA	Active_c1	0.9025	0.9315	318
	Active_c2	0.9001	0.9256	259
	Active_c3	0.8577	0.8899	275
	Active_c4	0.8315	0.8571	330
Ours	Active_r1	0.8945	0.9077	357
	Active_r2	0.9035	0.9048	384
	Active_r3	**0.9220**	**0.9643**	402
Active learning	Ablation_r1	0.8964	0.9196	421
	Ablation_r2	0.8969	0.9345	480
	Ablation_r3	0.9047	0.9405	512
	Ablation_r4	0.9053	0.9455	540
Random learning	Random_v1	0.8874	0.8989	402
	Random_v2	0.8537	0.8720	402

efforts. We carry out experiments using the same deep learning model. Specifically, this counterpart only employs active learning but not self-paced learning, also termed as ablation counter-part. Therefore, all most informative samples are selected in each round. In details, 201, 59, 32, 28 most informative sam-ples are annotated and added into the annotated samples in each round. In this way, the FROC curves of detection results on the test dataset are shown in Fig. 9, where the rounds of the ablation counterpart are termed Ablation_r1 to Ablation_r4. We can observe that the updated model on Ablation_r4 presents the best 0.9053 PAUC and 0.9455 TPR@2.0FPI, which inferior to our learning framework (i.e., 0.9220 PAUC and 0.9643 TPR@2.0FPI of Active_r3). Moreover, the ablation counterpart annotates 540 samples in total, which is more than that of our

learning frame-work (i.e., 402 samples). Therefore, the experimental results of the first counterpart suggest that our proposed learning frame-work achieves better performance and annotates lesser samples over the state-of-the-art counterpart.

4.5.2. Random learning (RL)

We employ random learning (RL) as another counterpart, which uses random query strategy to select unannotated samples. The selected samples are annotated and added into the initial annotated samples. Then, the extended annotated samples are used to update the initial model. With the same workload of annotation, we randomly select 182 samples from the initial unannotated samples for manual annotation. We conduct two group of experiments for random learning independently, termed

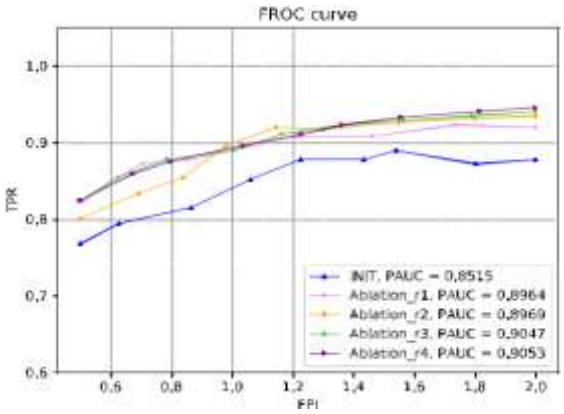


Fig. 9. Experimental results of the active learning.

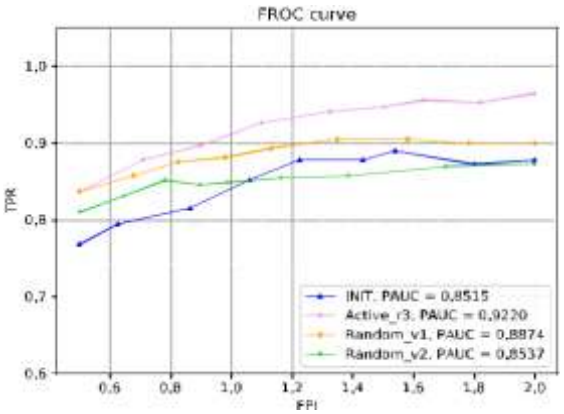


Fig. 10. Performance comparison between our learning framework and the random learning.

as Random_v1 and Random_v2. The FROC curves of detection results on the test dataset are shown in Fig. 10. We can observe that the Random_v1 and Random_v2 obtain 0.8874 PAUC, 0.8989 TPR@2.0FPI and 0.8537 PAUC, 0.8720 TPR@2.0FPI on the test dataset, respectively. Obviously, our learning framework significantly outperforms this counterpart. Additionally, the random query strategy is unstable and unreliable, because the Random_v1 and Random_v2 have distinct difference.

4.6. Discussion

Although we have already shown superiority of our proposed method in the experimental results, there are some technical essentials need to be discussed.

The average time of annotating a mammogram is approximate 5 to 10 min. Furthermore, the variation of annotations from different radiologists need more time to consult. In our proposed method, the radiologists only annotate 402 mammograms out of 1,912 mammograms in training dataset, about 20% of the total training dataset. In each round of active learning and self-paced learning, we train the FCN with 7,000 training iterations, about 4 h. The prediction time of unannotated samples is much less, e.g., 3 samples per second. As the results,

our method is converge in 4 rounds. Therefore, the time cost of our method is much less than that of manually annotating all training dataset.

In machine learning, the validation dataset is used to assist the training by evaluating its periodically in the training. Based on the evaluated result on the validation dataset, the training can be early stopped, or the model that achieves the best result on the validation dataset is selected. Therefore, the validation dataset is involved in the training process, and can restrain the overfitting. However, it is not mandatory in all machine learning tasks, espe-cially in the situation of indirect metrics for evaluation [49,50]. In our experiments, the FROC curves and the PAUCs are calculated from the heatmaps with unfixed multiple thresholds. In order to guarantee the FPI range of [0.5,2.0] in the FROC curves, we need manually adjust the upper bound and lower bound of thresholds, thus we cannot automatically calculated the FROC curves and PAUCs of the validation dataset in the training process periodi-cally. Therefore, we do not employ the validation dataset in our experiments, and we also do not use the test dataset to assist the training as the validation dataset.

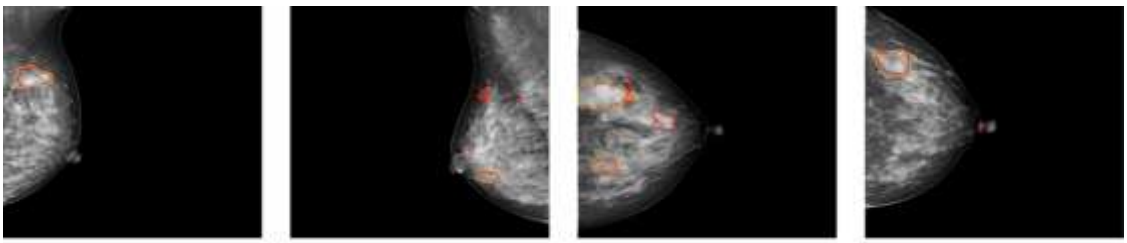
The k-fold cross-validation is very useful in the situation of the dataset with less samples [49]. However, we have a large number of samples in our dataset. To perform the k-fold cross-validation, it might have to annotate all samples. It is time-consuming and tedious. On the other hand, the main purpose of our proposed method is to minimize the number of manual annotation. Besides, the philosophy of self-paced learning provides robust learning process for FCN, thus our proposed method can present stable and reliable performance as the experimental results.

Accurately detecting the tiny objects is a challenge task in computer vision. Many popular object detection methods, such as Faster RCNN, SSD and YOLO, have this problem. For example, the Faster RCNN employs region proposal networks (RPN) to generate region proposals. The RPN locates the key points on the small-scale feature maps, and then uses multiple anchor boxes to generate region proposals on the original image. The region proposals are used for classification in subsequence, thus the anchor boxes cannot be too small. We employ FCN to make spatial density prediction (i.e., heatmap), then use the multiple thresholds to find out suspicious lesions. In order to achieve better segmentation contour, the FCN employ skip connections to acquire detailed information from low-level large-scale feature maps, which can enhance the recognition capability of tiny le-sions. Because the heatmap has same resolution with the original image, the FCN have higher detection sensitivity for tiny objects.

Mass embedded inside the dense tissues is the most difficult type (hard example), and the samples of this type are relatively less. In our dataset, parts of samples have this issue. Manual annotation for these samples is also difficult because the expe-rienced radiologists cannot find out the exact counters of the masses embedded inside the dense tissues. To address this issue, our proposed method employs focal loss to perform hard exam-ple mining. As our experimental results, the focal loss improves the performance. Besides, our proposed method employs active learning to preferentially select the informative samples (most of them are hard examples) for manual annotation, and employs self-paced learning to provide a robust learning process, i.e., grad-ually increase the number of hard examples at the latter round. Therefore, our proposed method is just right to address the issue of mass embedded inside the dense tissues. Some examples of dense cases are shown in Fig. 11.

V. CONCLUSION

This study develops a novel learning framework for breast mass detection that incorporated deep active learning (DAL) and self-paced learning (SPL) paradigm. The efficient learning strategy in our proposed learning framework can obtain better perfor-mance with minimum annotation efforts. Specifically, we employ focal loss in the deep learning model to tackle the class imbalance problem. In order to find out the most informative samples, we propose an informativeness query algorithm to rank the large amounts of unannotated samples. Then, we propose a self-paced



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

Fig. 11. Some examples of dense cases from our test dataset. The orange contours denote the manual annotations (ground truth) and the red contours denote the predicted suspicious mass lesions. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

sampling algorithm to select a number of the most informative samples for manual annotation. The experimental results suggest that our proposed learning framework achieves superior perfor-mance over the counterparts. Moreover, our proposed learning framework dramatically reduces the requirement of the anno-tated samples, i.e., about 20% of all training data. In the future work, we will investigate more measurements in the sample selection criteria, such as noise level. Besides, we would like to extend this methodology to other similar tasks in different imaging modalities, such as pathology and MRI.

Declaration of competing interest

The authors declare that there is no conflict of interest with other organizations or researchers.

REFERENCES

[1] American Cancer Society, Breast cancer facts & figures 2017–2018, 2017, https://www.cancer.org/research/cancer-facts-statistics/breast-cancer-facts-figures.html.
[2] B. Stewart, C.P. Wild, et al., World cancer report 2014, Health, 2017.
[3] A. Oliver, J. Freixenet, J. Marti, E. Perez, J. Pont, E.R. Denton, R. Zwiggelaar, A review of automatic mass detection and segmentation in mammographic images, Med. Image Anal. 14 (2) (2010) 87–110.
[4] J. Long, E. Shelhamer, T. Darrell, Fully convolutional networks for semantic segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 3431–3440.
[5] O. Ronneberger, P. Fischer, T. Brox, U-net: convolutional networks forbiomedical image segmentation, in: International Conference on Medical Image Computing and Computer-assisted Intervention, Springer, 2015, pp. 234–241.
[6] Y. Bengio, J. Louradour, R. Collobert, J. Weston, Curriculum learning, in: Proceedings of the 26th Annual International Conference on Machine Learning, ACM, 2009, pp. 41–48.
[7] M.P. Kumar, B. Packer, D. Koller, Self-paced learning for latent variable models, in: Advances in Neural Information Processing Systems, 2010, pp. 1189–1197.
[8] E. Sangineto, M. Nabi, D. Culibrk, N. Sebe, Self paced deep learning for weakly supervised object detection, 2016, arXiv preprint arXiv:1605.07651.
[9] D. Zhang, D. Meng, J. Han, Co-saliency detection via a self-paced multiple-instance learning framework, IEEE Trans. Pattern Anal. Mach. Intell. 39 (5) (2017) 865–878.
[10] L. Lin, K. Wang, D. Meng, W. Zuo, L. Zhang, Active self-paced learning for cost-effective and progressive face identification, IEEE Trans. Pattern Anal. Mach. Intell. 40 (1) (2018) 7–19.
[11] J. Arevalo, F.A. González, R. Ramos-Pollán, J.L. Oliveira, M.A.G. Lopez, Representation learning for mammography mass lesion classification with convolutional neural networks, Comput. Methods Programs Biomed. 127 (2016) 248–257.
[12] T. Kooi, G. Litjens, B. van Ginneken, A. Gubern-Mérida, C.I. Sánchez, R. Mann, A. den Heeten, N. Karssemeijer, Large scale deep learning for computer aided detection of mammographic lesions, Med. Image Anal. 35 (2017) 303–312.
[13] N. Dhungel, G. Carneiro, A.P. Bradley, A deep learning approach for the analysis of masses in mammograms with minimal user intervention, Med. Image Anal. 37 (2017) 114–128.
[14] D. Ribli, A. Horváth, Z. Unger, P. Pollner, I. Csabai, Detecting and classifying lesions in mammograms with deep learning, Sci. Rep. 8 (1) (2018) 4165.
[15] S. Ren, K. He, R. Girshick, J. Sun, Faster r-cnn: towards real-time object de-tection with region proposal networks, in: Advances in Neural Information Processing Systems, 2015, pp. 91–99.
[16] M.A. Al-masni, M.A. Al-antari, J.-M. Park, G. Gi, T.-Y. Kim, P. Rivera, E. Valarezo, M.-T. Choi, S.-M. Han, T.-S. Kim, Simultaneous detection and classification of breast masses in digital mammograms via a deep learning YOLO-based CAD system, Comput. Methods Programs Biomed. 157 (2018) 85–94.
[17] M.A. Al-antari, M.A. Al-masni, M.-T. Choi, S.-M. Han, T.-S. Kim, A fully in-tegrated computer-aided diagnosis system for digital x-ray mammograms via deep learning detection, segmentation, and classification, Int. J. Med. Inform. 117 (2018) 44–54.
[18] M.G. Ertosun, D.L. Rubin, Probabilistic visual search for masses within mammography images using deep learning, in: Bioinformatics and Biomedicine (BIBM), 2015 IEEE International Conference on, IEEE, 2015, pp. 1310–1315.
[19] A.A. Novikov, D. Lenis, D. Major, J. Hladuvka, M. Wimmer, K. Bühler, Fully convolutional architectures for multi-class segmentation in chest radiographs, IEEE Trans. Med. Imaging (2018).

[20] S.S.M. Salehi, D. Erdogmus, A. Gholipour, Auto-context convolutional neural network (auto-net) for brain extraction in magnetic resonance imaging, IEEE Trans. Med. Imag. 36 (11) (2017) 2319–2330.
[21] L. Chen, P. Bentley, K. Mori, K. Misawa, M. Fujiwara, D. Rueckert, DRINet for medical image segmentation, IEEE Trans. Med. Imaging (2018).
[22] G. Wang, W. Li, M.A. Zuluaga, R. Pratt, P.A. Patel, M. Aertsen, T. Doel, A.L. David, J. Deprest, S. Ourselin, et al., Interactive medical image segmentation using deep learning with image-specific fine tuning, IEEE Trans. Med. Imag. 37 (7) (2018) 1562–1573.
[23] Y. Li, L. Xu, J. Rao, L. Guo, Z. Yan, S. Jin, A y-net deep learning method for road segmentation using high-resolution visible remote sensing images, Remote Sens. Lett. 10 (4) (2019) 381–390.
[24] N. Roy, A. McCallum, Toward Optimal Active Learning Through Monte Carlo Estimation of Error Reduction, ICML, Williamstown, 2001, pp. 441–448.
[25] K. Brinker, Incorporating diversity in active learning with support vector machines, in: Proceedings of the 20th International Conference on Machine Learning, ICML-03, 2003, pp. 59–66.
[26] S. Dutt Jain, K. Grauman, Active image segmentation propagation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2864–2873.
[27] R. Wang, X.-Z. Wang, S. Kwong, C. Xu, Incorporating diversity and infor-mativeness in multiple-instance active learning, IEEE Trans. Fuzzy Syst. 25(6) (2017) 1460–1475.
[28] B. Zhang, Y. Wang, F. Chen, Multilabel image classification via high-order label correlation driven active learning, IEEE Trans. Image Process. 23 (3) (2014) 1430–1441.
[29] J. Yang, S. Li, W. Xu, Active learning for visual image classification method based on transfer learning, IEEE Access 6 (2018) 187–198.
[30] B. Liu, V. Ferrari, Active learning for human pose estimation, in: Proceed-ings of the IEEE International Conference on Computer Vision, 2017, pp. 4363–4372.
[31] C. Liu, L. He, Z. Li, J. Li, Feature-driven active learning for hyperspectral image classification, IEEE Trans. Geosci. Remote Sens. 56 (1) (2018) 341–354.
[32] J. Melendez, B. van Ginneken, P. Maduskar, R.H. Philipsen, H. Ayles, C.I. Sánchez, On combining multiple-instance learning and active learning for computer-aided detection of tuberculosis, IEEE Trans. Med. Imag. 35 (4) (2016) 1013–1024.
[33] H. Su, Z. Yin, S. Huh, T. Kanade, J. Zhu, Interactive cell segmentation based on active and semi-supervised learning, IEEE Trans. Med. Imag. 35 (3) (2016) 762–777.
[34] W. Shao, L. Sun, D. Zhang, Deep active learning for nucleus classification in pathology images, in: Biomedical Imaging (ISBI 2018), 2018 IEEE 15th International Symposium on, IEEE, 2018, pp. 199–202.
[35] J. Folmsbee, X. Liu, M. Brandwein-Weber, S. Doyle, Active deep learning: improved training efficiency of convolutional neural networks for tissue classification in oral cavity cancer, in: Biomedical Imaging (ISBI 2018), 2018 IEEE 15th International Symposium on, IEEE, 2018, pp. 770–773.
[36] J. Wu, S. Ruan, C. Lian, S. Mutic, M.A. Anastasio, H. Li, Active learning with noise modeling for medical image annotation, in: Biomedical Imaging (ISBI 2018), 2018 IEEE 15th International Symposium on, IEEE, 2018, pp. 298–301.
[37] H. Ranganathan, H. Venkateswara, S. Chakraborty, S. Panchanathan, Deep active learning for image classification, in: Image Processing (ICIP), 2017 IEEE International Conference on, IEEE, 2017, pp. 3934–3938.
[38] W. Zhao, Y. Kong, Z. Ding, Y. Fu, Deep active learning through cognitive information parcels, in: Proceedings of the 2017 ACM on Multimedia Conference, ACM, 2017, pp. 952–960.
[39] R. Girshick, Fast r-cnn, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 1440–1448.
[40] T.-Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollár, Focal loss for dense object detection, 2017, arXiv preprint arXiv:1708.02002.
[41] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, 2014, arXiv preprint arXiv:1409.1556.
[42] D.P. Kingma, J. Ba, Adam: a method for stochastic optimization, 2014, arXiv preprint arXiv:1412.6980.
[43] L. Jiang, D. Meng, Q. Zhao, S. Shan, A.G. Hauptmann, Self-paced curriculum learning, in: AAAI, vol. 2, 2015, p. 6.
[44] M. Heath, K. Bowyer, D. Kopans, R. Moore, P. Kegelmeyer, The digital database for screening mammography, in: Digital Mammography, 2000, pp. 431–434.
[45] I.C. Moreira, I. Amaral, I. Domingues, A. Cardoso, M.J. Cardoso, J.S. Cardoso, Inbreast: toward a full-field digital mammographic database, Acad. Radiol. 19 (2) (2012) 236–248.
[46] D.P. Chakraborty, A brief history of free-response receiver operating characteristic paradigm data analysis, Acad. Radiol. 20 (7) (2013) 915–919.

[47] F. Chollet, et al., Keras, 2015, https://github.com/keras-team/keras.
[48] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, et al., Tensorflow: a system for large-scale machine learning, in: OSDI, vol. 16, 2016, pp. 265–283.
[49] L. Fei-Fei, J. Justin, S. Yeung, Cs231n: convolutional neural networks for visual recognition, 2018, https://github.com/cs231n/cs231n.github.io.
[50] T. Shah, Train, validation and test sets, 2017, https://tarangshah.com/blog/ 2017-12-03/train-validation-and-test-sets/.