Weakly-Supervised Self-Training for Breast Cancer Localization*

Gongbo Liang[1], Xiaoqin Wang[2], Yu Zhang[1], Nathan Jacobs[1]

***Abstract*— The use of deep learning methods has dramatically increased the state-of-the-art performance in image object localization. However, commonly used supervised learning methods require large training datasets with pixel-level or bounding box annotations. Obtaining such fine-grained annotations is extremely costly, especially in the medical imaging domain. In this work, we propose a novel weakly supervised method for breast cancer localization. The essential advantage of our approach is that the model only requires image-level labels and uses a self-training strategy to refine the predicted localization in a step-wise manner. We evaluated our approach on a large, clinically relevant mammogram dataset. The results show that our model significantly improves performance compared to other methods trained similarly.**

***Index Terms*— Object localization, mammography, convolutional neural network**

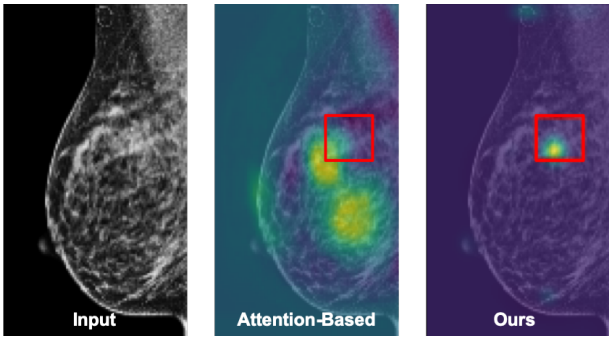

Fig. 1. Breast tumor localization example of a given input mammogram (left), the prediction of the attention-based method (middle), and the prediction of our method (right). The red box indicates the ground truth tumor localization. The heatmap shows the predictedu location.

I. Introduction

Recently, deep learning has demonstrated revolutionary potential in various medical imaging analysis tasks such as classification, localization, segmentation, image post-processing, treatment planning, etc [1], [2], [3], [4], [5], [6]. In object localization tasks, fully-supervised training methods usually require a large number of training images with bounding boxes (BBs) of region-of-interests (ROIs) or pixel-level annotations [7], [8], [9], [10]. However, such fine-grained annotations are usually not available for medical images, especially for clinically relative breast cancer dataset [11]. Obtaining the annotations usually is expensive and time-consuming because the annotator needs months or even years of professional training. In contrast to fully supervised training, weakly-supervised training uses coarser annotations, such as image-level labels [12], [13], [14], which can significantly reduce the time and cost for annotation.

Domain adaptation is one way to train an object localization network without fine-grained labels. This approach was proposed to deal with the scenarios in which a model trained on a source distribution (dataset) is used in the context of a different (but related) target distribution (dataset) [15]. Domain adaptation for weakly-supervised localization training shows promising results in natural imaging settings [16]. To use such a method, we need firstly to train a localization network on a source dataset with fine-grained labels. After the model is trained well, we can use domain adaptation to apply the pre-trained model on a different but related target dataset without requiring fine-grained labels. However, in the medical imaging domain, source datasets with fine-grained labels are usually not available in the real world.

Attention mechanism, which usually refers to trainable attention [17], [18], can be used for weakly-supervised object localization as well. An attention map highlights the important areas of a given image. Ideally, the highlighted areas should be the ROIs of a given image. We can use the attention map for object localization. However, in practice, not all of the important areas are necessary to be ROIs. Zhang et al. [19] proposed to use self-produced guidance (SPG) masks for object localization of natural images. A SPG mask is learned from an attention mask. Each pixel in the attention mask is labeled as one out of three classes using a thresholding method. Then, the SPG masks are used as auxiliary pixel-level supervision to facilitate the training of classification networks for object localization. Their method is the-state-of-art weakly-supervised localization performance on the ILSVRC [20] dataset.

Inspired by Zhang et al. [19], we propose to use the class activation mapping (CAM) mechanism and self-training strategies to train a tumor localization network using only the image-level labels. More specifically, we use CAM heatmaps to replace the attention maps in their work.

Class Activation Mapping was originally proposed for model decision visualization [21], [22], in which the pixel values of CAM heatmaps are associated with the contribution to the classification decision. A higher value indicates a higher contribution, which implies a higher possibility of the occurrence of the object-of-interest at that location. Unlike

*This study is sponsored by Grant No. IRG 16-182-28 from the American Cancer Society and Grant No. IIS-1553116 from the National Science Foundation. The experimental procedures involving human subjects described in this paper were approved by the Institutional Review Board.

[1]Gongbo Liang, Yu Zhang, and Nathan Jacobs are with Department of Computer Science, College of Engineering, University of Kentucky, USA {liang, yzh382, jacobs}@cs.uky.edu

[2]Xiaoqin Wang is with the Department of Radiology, College of Medicine, University of Kentucky, USA xiaoqin.wang@uky.edu

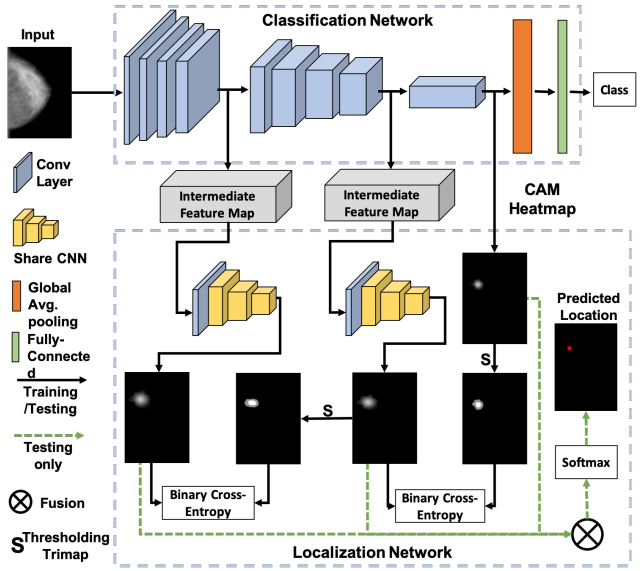

Fig. 2. An illustration of our weakly-supervised self-training breast cancer localization model: 1) an input image passes through the classification network to extract the intermediate feature maps and CAM heatmap; 2) the localization network is trained using a self-training strategy with the intermediate feature maps and CAM heatmap; 3) at the testing stage, softmax is applied on the fused CAM and localization network outputs to find the final object location.

attention maps, CAM heatmaps are only highlighting the most discriminative regions. The ROIs are usually much smaller in the medical imaging domain, and the ratio of image size to ROI size is often much higher, comparing with the natural imaging domain. For instance, a typical full-field digital mammogram (FFDM) size is 3328×4096 pixels. However, the size of a breast tumor could be as small as 10 pixels in diameter. CAM-based methods provide a more precise localization result than the attention-based methods (Figure 1).

We evaluated the proposed approach on a large, clinically relevant mammogram dataset, which was recently collected from a comprehensive breast care center. Our experiment results show that the proposed method significantly improves the performance of weakly-supervised breast cancer localization tasks.

II. Architecture

A. Network Overview

Given an input image at the training stage, a CAM heatmap can be learned from a classification network. A trimap (a pixel-level annotation for each pixel in a CAM heatmap, in which each pixel belongs to one of three classes in the trimap) can be derived from the CAM heatmap, which highlights the high confident foreground (ROI/tumor) pixels, the high confident background (non-ROI/non-tumor) pixels, and the unknown pixels. Then, the trimap can be used as the pseudo-pixel-level label in a self-training convolutional neural network (CNN) localization model (Figure 2). More specifically, we use the foreground and background pixels in the trimap as the pseudo-pixel-level label and use the corresponding areas in an intermediate feature map (the output from a higher convolutional layer) as the input to train a CNN model for the pixel-level labeling task. The prediction of this CNN can be used to generate another pseudo-pixel-level label (trimap) to train a new CNN model that takes another intermediate feature map from an even higher convolutional layer (Conv-layer) as the input. This self-training strategy can be repeated up to K times (K equals to the number of Conv-layers in the classification model).

At the testing stage, the predictions of all the self-trained CNN models were combined with the CAM heatmap. The softmax function will be applied to find the final predicted ROI localization.

B. CAM Heatmap Generalization

Class activation maps (CAM heatmaps) is generated using the global average pooling (GAP) in CNN classification networks. A CAM heatmap for a particular category indicates the discriminative image regions used by the CNN model to identify that category. More specifically, we first need to train a classification network with a GAP layer. The GAP layer follows the last Conv-layer in the network. After the GAP layer, we will have a fully-connected network follows by A softmax layer, which provides the classification decision of a given image. To generate the CAM heatmap of the predicted class, we need to: 1) get all the weights connected between the fully-connected layer and the softmax class of which we want to predict. If n feature maps are presented before the GAP layer, n weights will be received. 2) We compute the weighted sum of the n feature maps that come from the last Conv-layer. The weighted sum generates a heatmap of a particular class. The size of the heatmap is the same as the feature map. Please see [21] for more details.

C. Self-Training

Self-training of a localization network includes two components: a pseudo-label generating strategy and a CNN model trained with the fully supervised training style. In our study, we use the self-training strategy to train multiple CNN models recursively. Each CNN model takes the output of a Conv-layer as the input and predicts a heatmap. The heatmap indicates the probability of being a tumor for each specific pixel in the input image. The pseudo-label used in the training of the base CNN model (the first model in the recursive sequence) is derived from the CAM heatmap using a thresholding method. The prediction of the base CNN model is used to generate the pseudo-label for the next CNN model, which trains in the same fashion.

More specifically, given an input image, I, we first feed it to a classification network and extract the CAM heatmap, C, and multiple intermediate feature maps, $\{F_i\}$, where $i \leq K$, K equals the number of Conv-layers in the classification model. We generate a trimap, M_C, of C using two thresholds t_f and t_b. For each pixel, p_j in C, if $p_j > t_f$, p_j is labeled as foreground; if $p_j < t_b$, p_j is labeled as background; if $t_b \leq p_j \leq t_f$, p_j is labeled as unknown. The foreground

and background pixels in M_C are used to train a base CNN model (CNN_{base}).

The CNN_{base} takes F_i as the input and predicts the pixel-level label for each pixel in I. The predictions form a new heatmap, M'_C. Ideally, the high confidence foreground and background areas in M'_C and M_C should be identical to each other. The CNN_{base} also predicts binary pixel labels of the area that was signed as unknown in M_C. A new trimap, M_{K-1}, is derived from M'_C using the same thresholding method. M_{K-1} is used to train a new CNN model, CNN_{K-1}, which takes F_{i-1} as the input and predicts M'_{K-1}. We repeat this process recursively until CNN_1 is trained, which uses F_1 as the input and M_1 as the ground truth label.

Binary cross-entropy (Equation 1) loss is used in all the CNN models.

$$BEC = -\frac{1}{N}\sum_{i=1}^{N} y_i \cdot \log(p(\hat{y}_i)) + (1-y_i)\cdot\log(1-p(\hat{y}_i)), \quad (1)$$

where y is the label, $p(\hat{y})$ is the predicted probability of the given data point, and N is the number of data points.

D. Implementation

We used Inception-V3 as our classification network in this study. We removed all the layers after the last Inception block. Then, we added two Conv-layers of kernel size 3×3, stride 1 with 1024 kernels, a global average pooling layer, a fully-connected layer, and a softmax layer. We used the output of the last Conv-layer to compute the CAM heatmap. We extracted the outputs of the third and eighth Inception blocks as the intermediate feature maps.

The localization network contained two CNN models, one for each intermediate feature map. Each of the CNN models had three Conv-layers, followed by a sigmoid layer. The first Conv-layers of the two CNN models had 288 and 768 3×3 kernels, respectively. The second Conv-layers of both CNNs contained 512 1×1 kernels, and the third Conv-layers of both CNNs had one 1×1 kernel. The weights of the second and third layers were shared between the two CNNs.

The model was implemented in PyTorch and trained with batch size 8. The initial learning rate was 0.001. SGD optimizer with a momentum of 0.9 was used in training. We chose the thresholds $t_f = 0.6$ and $t_b = 0.1$. We trained and tested the network on an Nvidia GTX 1080 GPU card with 8GB of memory.

III. Experiments

A. Dataset

We use the UKy dataset (a large, clinically related mammogram dataset) for this study. The dataset contains FFDM images for 779 positive cases and 3018 negative cases. All the mammography data were retrospectively collected from patients seen at a comprehensive breast imaging center in the United States from Jan 2014 to Dec 2017. All patients had mammograms in either craniocaudal (CC) view, mediolateral oblique (MLO) view, or both. Each image was reviewed by specialized breast radiologists. All the positive cases were proved with biopsy, and the negative cases were confirmed with more than two years of follow-up. The dataset contains cases with co-existing conditions, such as a prior benign biopsy and surgery.

The images also contain common foreign bodies, such as clips, markers, and pacemakers. The images were acquired with Hologic devices in 12-bit DICOM format at the resolution of 3328×4096 and downsampled to 832×832. Data augmentation was applied to all the positive images through a combination of reflection and rotation. Each original image was flipped horizontally and rotated by each of 90, 180, and 270 degrees. In total, 4175 positive images and 12072 negative images are used in the training stage. The dataset is randomly split into the training and validation sets on the patient-level with a $4:1$ ratio.

The training and validation sets were used for the classification model training. We manually annotated the ROIs of additional 138 positive images with bounding boxes for testing. These images were held out during the training stage and only used in the testing stage.

TABLE I
Localization performances.

Model	STL	FCN_{WSL}	$Ours$
Loc. AP	0.43	0.26	**0.52**

B. Result

We used STL [23] and FCN_{WSL} [24] as the baseline models in this study. STL was specifically designed for breast cancer localization. FCN_{WSL} was an extension of [12] on medical related tasks. The CAM-based weakly-supervised training methods were used in both models.

We evaluated our model on localization AP, which has been widely used in weakly-supervised object localization tasks [12], [13], [24], [23]. We calculated localization AP in the following way: if the predicted location lies within the ground truth bounding box of the same class or within a tolerance distance (d), the example is considered as true positive; otherwise, it is a false positive prediction. In our experiment, only the positive class is considered for localization AP since there is no ROI on the negative class. We chose d to be equal to 12 pixels, which is the mean of [24], [23].

Table I shows localization AP for the three models. Our model achieves 0.52 localization AP, which surpasses STL by 20.93% (0.43 localization AP). FCN_{WSL} only achieved 0.26 localization AP, which is only 50% of our model.

Figure 3 shows the prediction results of our model. The testing images are on the left, and the predicted heatmaps are on the right. The red boxes are the ground truth bounding boxes of the malignant tumors, which indicates the ground truth localization. We used the center pixel of each heatmap as the final predicted localization. If the pixel lies in the

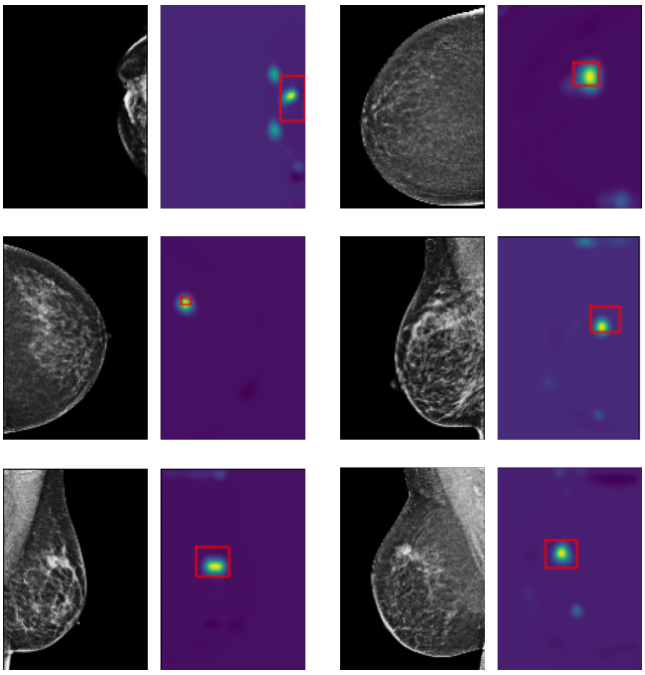

Fig. 3. Breast tumor localization examples generated with our method. The red boxes are the ground truth bounding boxes. The heatmaps show the predicted locations.

ground truth bounding boxes or within 12 pixels, we consider the prediction as true positive.

The figure shows that our model is able to predict correct locations for both mass and calcification cases. The figure also demonstrates that our model has the ability to work in very challenging cases, such as cases with prior surgery history (the top left example in the figure).

IV. Conclusion

We proposed a novel weakly-supervised breast cancer localization network. The proposed method only requires the image-level labels for training. No fine-grained annotation, such as bounding boxes or pixel-level labels, are needed in the training process. The model uses CAM to generate a pseudo-pixel-level label to train a localization network gradually in a self-training fashion. The evaluation result on a large clinically relevant mammogram dataset shows the proposed method has significantly improved the performance in object localization. We believe the proposed model is not only limited to breast tumor localization. It should be easily transferred to other medical imaging localization tasks with minor changes. We believe this work will serve as a strong baseline for future researchers.

References

[1] Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun, "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, vol. 542, no. 7639, pp. 115, 2017.

[2] Radu Paul Mihail, Gongbo Liang, and Nathan Jacobs, "Automatic hand skeletal shape estimation from radiographs," *IEEE Transactions on NanoBioscience*, vol. 18, no. 3, pp. 296–305, July 2019.

[3] Berkman Sahiner, Aria Pezeshk, Lubomir M Hadjiiski, Xiaosong Wang, Karen Drukker, Kenny H Cha, Ronald M Summers, and Maryellen L Giger, "Deep learning in medical imaging and radiation therapy," *Medical physics*, vol. 46, no. 1, pp. e1–e36, 2019.

[4] Yu Zhang, Xiaoqin Wang, Hunter Blanton, Gongbo Liang, Xin Xing, and Nathan Jacobs, "2d convolutional neural networks for 3d digital breast tomosynthesis classification," in *BIBM*, 2019, pp. 1013–1017.

[5] Qingsong Yang, Pingkun Yan, Yanbo Zhang, Hengyong Yu, Yongyi Shi, Xuanqin Mou, Mannudeep K Kalra, Yi Zhang, Ling Sun, and Ge Wang, "Low-dose ct image denoising using a generative adversarial network with wasserstein distance and perceptual loss," *IEEE transactions on medical imaging*, vol. 37, no. 6, pp. 1348–1357, 2018.

[6] Gongbo Liang, Xiaoqin Wang, Yu Zhang, Xin Xing, Hunter Blanton, Tawfiq Salem, and Nathan Jacobs, "Joint 2d-3d breast cancer classification," in *BIBM*, 2019.

[7] Ross Girshick, "Fast r-cnn," in *ICCV*, 2015, pp. 1440–1448.

[8] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi, "You only look once: Unified, real-time object detection," in *CVPR*, 2016, pp. 779–788.

[9] Dezső Ribli, Anna Horváth, Zsuzsa Unger, Péter Pollner, and István Csabai, "Detecting and classifying lesions in mammograms with deep learning," *Scientific reports*, vol. 8, no. 1, pp. 4165, 2018.

[10] Wentao Zhu, Chaochun Liu, Wei Fan, and Xiaohui Xie, "Deeplung: Deep 3d dual path nets for automated pulmonary nodule detection and classification," in *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2018, pp. 673–681.

[11] Xiaoqin Wang, Gongbo Liang, Yu Zhang, Hunter Blanton, Zachary Bessinger, and Nathan Jacobs, "Inconsistent performance of deep learning models on mammogram classification," *Journal of the American College of Radiology*, 2020.

[12] Thibaut Durand, Taylor Mordan, Nicolas Thome, and Matthieu Cord, "Wildcat: Weakly supervised learning of deep convnets for image classification, pointwise localization and segmentation," in *CVPR*, 2017, pp. 642–651.

[13] Maxime Oquab, Léon Bottou, Ivan Laptev, and Josef Sivic, "Is object localization for free?-weakly-supervised learning with convolutional neural networks," in *CVPR*, 2015, pp. 685–694.

[14] Hakan Bilen and Andrea Vedaldi, "Weakly supervised deep detection networks," in *CVPR*, 2016, pp. 2846–2854.

[15] Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan, "A theory of learning from different domains," *Machine learning*, vol. 79, no. 1-2, pp. 151–175, 2010.

[16] Taekyung Kim, Minki Jeong, Seunghyeon Kim, Seokeon Choi, and Changick Kim, "Diversify and match: A domain adaptive representation learning paradigm for object detection," in *CVPR*, 2019, pp. 12456–12465.

[17] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[18] Saumya Jetley, Nicholas A Lord, Namhoon Lee, and Philip HS Torr, "Learn to pay attention," *arXiv preprint arXiv:1804.02391*, 2018.

[19] Xiaolin Zhang, Yunchao Wei, Guoliang Kang, Yi Yang, and Thomas Huang, "Self-produced guidance for weakly-supervised object localization," in *ECCV*, 2018, pp. 597–613.

[20] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al., "Imagenet large scale visual recognition challenge," *IJCV*, vol. 115, no. 3, pp. 211–252, 2015.

[21] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba, "Learning deep features for discriminative localization," in *CVPR*, 2016, pp. 2921–2929.

[22] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *ICCV*, 2017, pp. 618–626.

[23] Sangheum Hwang and Hyo-Eun Kim, "Self-transfer learning for weakly supervised lesion localization," in *MICCAI*. Springer, 2016, pp. 239–246.

[24] Armine Vardazaryan, Didier Mutter, Jacques Marescaux, and Nicolas Padoy, "Weakly-supervised learning for tool localization in laparoscopic videos," in *Intravascular Imaging and Computer Assisted Stenting and Large-Scale Annotation of Biomedical Data and Expert Label Synthesis*, pp. 169–179. Springer, 2018.