LEARNING FROM WEAKLY-LABELED CLINICAL DATA FOR AUTOMATIC THYROID NODULE CLASSIFICATION IN ULTRASOUND IMAGES

Jianxiong Wang[1] *Shuai Li*[1] *Wenfeng Song*[1] *Hong Qin*[2] *Bo Zhang*[3] *Aimin Hao*[1]

[1] State Key Laboratory of Virtual Reality Technology and Systems, Beihang University, Beijing, China
[2] Department of Computer Science, Stony Brook University, Stony Brook, USA
[3] Department of Ultrasound, Chinese Academy of Medical Science &
Peking Union Medical College Hospital, Beijing, China

ABSTRACT

This paper proposes a semi-supervised learning method based on weakly-labeled data to automatically classify ultrasound (US) thyroid nodules. Key to our new approach is the unification of multi-instance learning (MIL) with deep learning. Benefiting from that, our method can directly use off-the-shelf clinical data, which involves no labels to indicate nodule classes. To this end, we take the US images of a patient as a bag, and take the corresponding pathology report as the bag label. Specifically, we first propose a bag generating method, wherein the detected thyroid nodules are considered as instances corresponding to certain bag. After that, we design an effective EM algorithm to train a convolutional neural network (CNN) for nodule classification. We conduct extensive experiments and comprehensive evaluations on different datasets, and all the experiments confirm that, our method significantly outperforms state-of-the-art MIL algorithms, which exhibits great potential in clinical applications.

Index Terms— Weakly-labeled data, Multi-instance learning (MIL), Convolutional neural network (CNN), Thyroid ultrasound image, Automatic nodule classification

1. INTRODUCTION AND PRIOR WORK

Nodular lesions of thyroid are very common among the general population [1]. According to National Cancer Institute[1], in 2017, it is estimated that 2,010 people will die of thyroid cancer and there will be 56,870 new disease cases, and this number continues to rise in recent years. At the imaging front, US imaging has been a dominant and preferred screening modality towards the clinical diagnosis of thyroid nodules [2] thanks to its sensitivity and convenience. In today's clinical practice, senior practitioners could pinpoint nodules by analyzing global context features, local geometry structure, and intensity variations, which would require rich clinical experiences accumulated from hundreds and thousands of nodule case studies.

To alleviate doctors' tremendous labor in the diagnosis procedure, recently many methods have been proposed to automatically classify thyroid nodules in US images, which can be roughly classified into two main categories: supervised methods and semi-supervised methods. For example, Chang et al. adopt SVM to select significant textural features and classify the nodular lesions [1]. In [3], artificial neural network and SVM are employed for the classification task, utilizing feature vectors derived from gray level co-occurrence (GLCM) features. Nugroho et al. use texture analysis to extract feature and employ MLP to classify cystic nodule from solid nodule [4]. In [5], GoogLeNet is fine-tuned to extract superior features, which are sent to a cost-sensitive random forest classifier to classify the images into "malignant" and "benign" cases. Liu et al. use a CNN model to generate semantic features and combine those features with HOG and LBP together to form a hybrid feature space. After that, a positive-sample-first majority voting and a feature-selection based strategy are employed for classification [6]. The above-mentioned methods all belong to the supervised learning paradigm, which means they all need accurately-annotated data. However, high-quality manual annotation is labor-intensive and time-consuming to obtain [7], especially in the field of medical images.

As for the semi-supervised methods, MIL [8], which can ease the burden of manual annotation naturally, has recently been used for thyroid nodule classification. In [9], the thyroid B-mode US image and the elastogram are regarded as a bag, of which local features of the B-mode image and global features of the elastogram are considered as instances of the bag, and SVM is employed to classify the lesion. Besides, MIL technique has also been used in the classification problems of gastric cancer with dual-energy CT imaging [10], breast US image [11] and histopathology cancer image [7]. These methods usually take single labeled image as a bag, in sharp contrast, our method can directly use the clinical data that is usually off-the-shelf in hospitals, and thus can further reduce the burden of manual annotation.

The salient contributions of this paper can be summarized

[1] Information at: https://seer.cancer.gov/statfacts/html/thyro.html

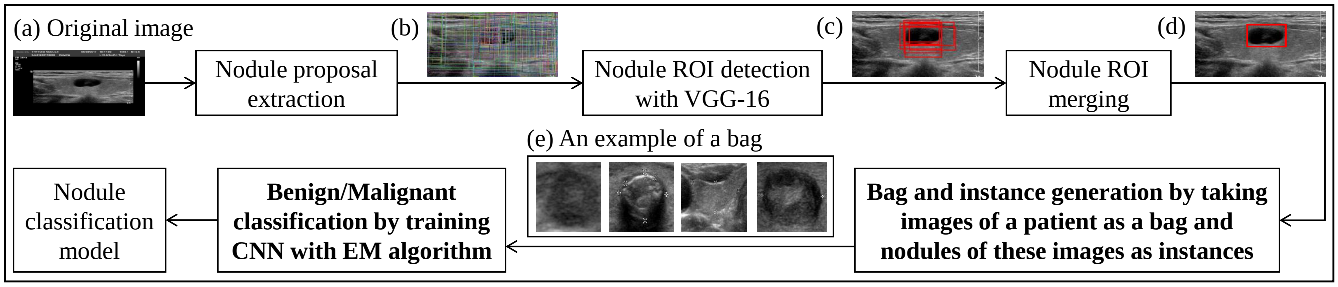
(a) Original image
Nodule proposal extraction
(b)
Nodule ROI detection with VGG-16
(c)
Nodule ROI merging
(d)
(e) An example of a bag
Bag and instance generation by taking images of a patient as a bag and nodules of these images as instances
Benign/Malignant classification by training CNN with EM algorithm
Nodule classification model

Fig. 1. The flowchart of our method.

as follows: (1) We propose a novel semi-supervised learning method for automatic US thyroid nodule classification, which does not need patch/image-level labels, and thus can significantly ease the burden of manually annotating training data; (2) We propose a novel EM algorithm to train CNN, which can solve multi-instance problems effectively and help improve the classification performance.

2. METHOD OVERVIEW

In this paper, we classify thyroid nodules in US images as benign or malignant based on clinical data, which usually involves US images, US reports and pathology reports, by treating patient-specific US images as a bag and the corresponding pathology report result as the bag label. If the bag label is positive, namely malignant, it means that there is at least one malignant nodule in the US images of that bag, and all nodules are benign otherwise.

The main steps of our proposed method are illustrated in Fig. 1. We first generate bag instances. Concretely speaking, original images are first pre-processed to remove useless regions. On that basis, we generate thousands of nodule proposals with a proposal extraction method. After that, we classify the proposals as nodules or non-nodules using a pre-trained nodule/non-nodule classification VGG-16 [12] to obtain a few nodule ROIs (regions of interest) with high recall rate. A simply-modified non-maximum suppression (NMS) algorithm is subsequently used to merge and regenerate nodule ROIs to get the final nodules. These nodules are subsequently considered as instances corresponding to certain bags. And then we train a CNN based nodule classification model with our carefully-designed EM algorithm on weakly-labeled training data. In the test phase, we feed the thyroid nodules in US images to our trained nodule classification model to obtain benign or malignant classification results.

3. NODULE INSTANCE GENERATION OF MIL BAG

3.1. Potential US Nodule Proposal Extraction

Original US images usually involves some useless regions around the image's edges (see Fig. 1(a)), which would meaninglessly increase the computational burden. Thus, we first remove those useless regions with traditional image processing methods.

After that, we employ a proposal extraction method to locate the potential nodule regions (see Fig. 1(b)). We have compared three excellent proposal extraction methods, including edge boxes [13], selective search [14] and BING [15] with sufficient experiments, and the method of edge boxes is proved to be the best choice because of its accuracy and efficiency. For each image, we extract at most 10,000 proposals with edge boxes ($\alpha = 0.7, \beta = 0.75$). Here both α and β are the parameters involved in edge boxes, which control the density of the potential nodule proposals.

3.2. Nodule ROI Detection

We train VGG-16 for nodule ROI detection with 3459 US images from scratch. Each thyroid nodule is framed with a rectangle box by radiologists. Although these images are manually labeled, radiologists only need to find the nodules, which is much easier and faster than accurately assigning benign or malignant label to each nodule.

The positive examples consist of all ground truth nodules and those proposals having more than 75% overlapping area with respect to the ground truth. The negative examples are selected from the proposals generated by edge boxes, which should have less than 30% overlapping area with respect to certain positive example. To avoid near-duplicate negative examples, a negative example is excluded if it has more than 50% overlapping area with another negative example, and we select at most 100 negative examples from each image according to the descending order of their overlapping values with positive examples. Here all images are resized to 64×64.

Since positive examples are much less than negative ones, we apply data augmentation approaches to generate new positive examples. Specifically, all positive examples are first resized to 75×75, and then 9 randomly-cropped 64×64 patches from these images are added to positive examples. Meanwhile, during training we also employ some common data augmentation techniques such as rotation, shift, and flip.

We randomly select 3000 positive and 3000 negative examples as validation data and train VGG-16 for 100 epochs. The initial learning rate is 0.001 and decays 0.05 every two epochs and the momentum is 0.9. As a result, we save VGG-16 when training epoch is 95 due to the smallest validation loss. We take the regions detected by the pre-trained VGG-16 as nodule ROIs (see Fig. 1(c)).

3.3. Nodule Instance Generation based on ROI Merging

In practice, we usually get more than one ROIs belonging to the same object. One indispensable component, named NMS, is widely used as a post-processing algorithm responsible for merging these ROIs. When directly applying the original NMS algorithm, we find that the refined nodule ROIs in some images are not as good as we expected due to the fact that NMS tends to remove a good ROI if the score of that ROI is lower than another ROI. Thus, we think it is better to generate new ROIs.

We propose a simple yet effective solution by averaging the positions of existing ROIs whose IoU value with each ROI merged by original NMS is smaller than a NMS threshold (see Fig. 1(d)). We set the NMS threshold to be 0.3, which is empirically determined. By regenerating, the IoU value between the ground truth and our regenerated ROI has an average increase of 0.11, that is 0.7536 and 0.7073 for benign and malignant images, respectively. The final nodule detection rate for benign and malignant images are 98.18% and 94.32%, respectively. And the average counts of final nodules in benign and malignant images are 2.83 and 2.69, respectively.

So far, we consider all the nodules as instances of the corresponding bags (see Fig. 1(e)). If the label is positive, namely malignant, it means there is at least one of the nodules in that bag is malignant, and all nodules are benign otherwise.

4. BENIGN/MALIGNANT CLASSIFICATION BASED ON SEMI-SUPERVISED DEEP LEARNING

Based on the generated nodule instances, though MIL method can be used for such weakly-supervised problem naturally, we integrate MIL idea with more powerful CNN to further improve the performance of MIL method. In fact, original CNNs belong to supervised learning paradigm that needs a large amount of accurately annotated data. To this end, we propose a novel yet effective EM algorithm to enable CNN to learn from weakly labeled data, and experimental results demonstrate that our method significantly outperforms several state-of-the-art MIL algorithms.

Algorithm 1 Our EM algorithm to train CNN

```
Set epo = 0, MAX_EPOCH = 50, MIN_LOSS = 0.01
Train only fully-connected layer of CNN by treating all
    instances in positive bags as positive samples and all
    instances in negative bags as negative samples
Compute $L^v$ (loss on validation set) with CNN
while epo<MAX_EPOCH and $L^v$>MIN_LOSS do
    E step:
    Compute the probability of each instance to be positive in positive bags with CNN
    Treat certain percentage instances with the highest
        probabilities in positive bags as positive samples and all
        instances in negative bags as negative samples
    M step:
    Train CNN for one epoch
    Compute $L^v$ with CNN
    epo = epo + 1
end while
```

The details of our algorithm are documented in Algorithm 1. The primary idea of our method is that, we assume at least certain percentage of the instances in positive bags are positive, and we train CNN iteratively under the assumption with a novel EM algorithm. Specifically, in E step, we compute the probabilities of the instances in positive bags to be positive, and set certain percentage of the instances with the highest probabilities in positive bags to have positive label. In M step, we train CNN with examples that satisfy: positive examples comprise positive instances in positive bags, and negative samples comprise all the instances in negative bags. To get good positive examples at the beginning, we train the fully-connected layer only by treating all instances in positive bags as positive samples and all instances in negative bags as negative samples. After that, we train a CNN based nodule classification model with our EM algorithm till the validation loss is no greater than 0.01 within at most 50 epochs. We use AlexNet [16] as a basic CNN model in this paper and the percentage of positive instances in positive bags is empirically determined, which will be discussed in Section 5.2.

5. EXPERIMENTS AND EVALUATIONS

5.1. Training and Test Datasets

We have two datasets. A dataset provided by Peking Union Medical College Hospital (PUMCH dataset), consists of clinical data for training and 800 thyroid nodule images (400 benign and 400 malignant) for validation and test. The clinical data contains 489 malignant cases and 130 benign cases. Another dataset, to our best knowledge, is the only publicly available database (Open database), of which 70 nodules are labeled as benign and 124 are malignant [17].

5.2. Evaluations on PUMCH Dataset

In our experiments, the percentage value, indicating at least how many instances in positive bags are positive, is chosen from the set {0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7} empirically. We first randomly choose 200 benign and 200 malignant images as a validation set and the remaining as a test set. Then, we train our nodule classification model with Algorithm 1 on

Table 1. Classification performance comparison among different methods on PUMCH test set.

Methods	Sensitivity	Specificity	Accuracy
Our Method	**0.9000**	**0.8650**	**0.8825**
miFV (GLCM)	0.3700	0.9250	0.6475
miFV (SBN)	0.5200	0.9500	0.7350
MIBoosting (GLCM)	0.8900	0.6700	0.7800
MIBoosting (SBN)	0.8600	0.5650	0.7125

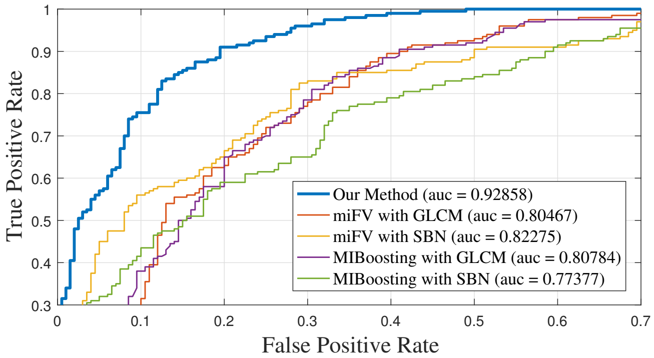

Fig. 2. ROC curves of different methods on PUMCH test set.

each value and compute validation loss. Finally, we choose the value that can result in the smallest validation loss. The initial learning rate is 0.0001, which decays 0.1 after each epoch, and the momentum is 0.9. For data augmentation, all images are first resized to 256 × 256, and then we randomly crop 227 × 227 patches as input during training. As a result, we choose the percentage value as 0.5 due to the smallest validation loss.

We conduct the classification performance comparison on the PUMCH test set between our method and two other excellent MIL algorithms (MIBoosting [18] and miFV [19]) according to an empirical study [20]. We extract SBN feature that is recommended in [20] and GLCM feature that is widely used in US image classification for these two algorithms. The implementation of MIBoosting is based on WEKA [21] and miFV from [19]. In MIBoosting, we use the pruned "J48" as the base learner and the maximum number of the boost iterations is chosen from $\{10, 50, 100\}$ empirically. In miFV, the center number of GMM is chosen from $\{1, 2, 3, 4, 5, 7, 10\}$ empirically, and the PCA energy is 1.0. We only report the best results of these two algorithms, which are chosen from the results obtained under different parameter values of these two algorithms.

The classification performance comparison among different methods on PUMCH test set are documented in Table 1, and the corresponding ROC curves are shown in Fig. 2. We can see that, our method significantly outperforms the other two state-of-the-art MIL algorithms. In Table 2, we also document the performance of our method for varying-size nodules' classification over PUMCH test set.

Table 2. Performance of our method for varying-size nodules' classification on PUMCH test set.

Nodule Size	Count	Sensitivity	Specificity	Accuracy
<1 cm	190	0.8703	0.9559	0.9316
1-3 cm	197	0.9058	0.6441	0.8274
3-5 cm	12	1.0000	1.0000	1.0000
>5 cm	1	—	1.0000	1.0000

5.3. Evaluations on Open Database

We evaluate the generalization performance of our method on an open database. We randomly choose 60 benign images and 80 malignant images to fine-tune our nodule classification model and 10 benign images and 44 malignant images for test, respectively. For fair comparison, we also fine-tune the original AlexNet. The initial learning rate is 0.00005, which decays 0.05 after each epoch, and the momentum is 0.9. The data augmentation techniques in Section 5.2 are also used here. The classification performance on the test set of open database is documented in Table 3. As can be seen, our nodule classification model significantly outperforms the original AlexNet, which manifests the effectiveness of our method.

Table 3. Classification performance comparison between our model and the original AlexNet on test set of open database by fine-tuning.

Methods	Sensitivity	Specificity	Accuracy
Our model	**0.8182**	**0.8000**	**0.8091**
AlexNet	0.7045	0.6000	0.6523

6. CONCLUSION

In this paper, we have advocated a novel semi-supervised method for automatic thyroid nodule classification in ultrasound images. We also propose a novel EM algorithm, with which traditional CNN can be extended to solve weakly-labeled learning problems effectively. Experimental results demonstrate that our method has superior advantages over the state-of-the-art MIL algorithms, which suggests its great potential for the clinical applications of US-based smart thyroid nodules diagnosis. Next, more comprehensive user studies and evaluations about equipping our method's framework with other popular networks deserve our immediate efforts.

7. ACKNOWLEDGMENTS

This research is supported in part by National Natural Science Foundation of China (NO. 61672077 and 61532002), Applied Basic Research Program of Qingdao (NO. 161013xx), National Science Foundation of USA (NO. IIS-0949467, IIS-1047715, IIS-1715985, and IIS-1049448), National Key R&D Program of China (NO. 2017YFF0106407), and capital health research and development of special 2016-1-4011.

8. REFERENCES

[1] C.-Y. Chang, S.-J. Chen, and M.-F. Tsai, "Application of support-vector-machine-based method for feature selection and classification of thyroid nodules in ultrasound images," *Pattern Recognition*, vol. 43, no. 10, pp. 3494–3506, 2010.

[2] U.R. Acharya, G. Swapna, S.V. Sree, F. Molinari, S. Gupta, R.H. Bardales, A. Witkowska, and J.S. Suri, "A review on ultrasound-based thyroid cancer tissue characterization and automated classification," *Technology in cancer research & treatment*, vol. 13, no. 4, pp. 289–301, 2014.

[3] H. Gireesha and S. Nanda, "Thyroid nodule segmentation and classification in ultrasound images," *International Journal of Engineering Research and Technology*, 2014.

[4] H.A. Nugroho, M. Rahmawaty, Y. Triyani, and I. Ardiyanto, "Texture analysis for classification of thyroid ultrasound images," in *IES*. IEEE, 2016, pp. 476–480.

[5] J. Chi, E. Walia, P. Babyn, J. Wang, G. Groot, and M. Eramian, "Thyroid nodule classification in ultrasound images by fine-tuning deep convolutional neural network," *Journal of digital imaging*, vol. 30, no. 4, pp. 477–486, 2017.

[6] T. Liu, S. Xie, J. Yu, L. Niu, and W. Sun, "Classification of thyroid nodules in ultrasound images using deep model based transfer learning and hybrid features," in *ICASSP*. IEEE, 2017, pp. 919–923.

[7] Y. Xu, J.-Y. Zhu, E. Chang, and Z. Tu, "Multiple clustered instance learning for histopathology cancer image classification, segmentation and clustering," in *CVPR*. IEEE, 2012, pp. 964–971.

[8] T.G. Dietterich, R.H. Lathrop, and T. Lozano-Pérez, "Solving the multiple instance problem with axis-parallel rectangles," *Artificial Intelligence*, vol. 89, no. 1, pp. 31–71, 1997.

[9] J. Ding, H. Cheng, J. Huang, and Y. Zhang, "Multiple-instance learning with global and local features for thyroid ultrasound image classification," in *BMEI*. IEEE, 2014, pp. 66–70.

[10] C. Li, C. Shi, H. Zhang, Y. Chen, and S. Zhang, "Multiple instance learning for computer aided detection and diagnosis of gastric cancer with dual-energy ct imaging," *Journal of biomedical informatics*, vol. 57, pp. 358–368, 2015.

[11] J. Ding, H. Cheng, J. Huang, J. Liu, and Y. Zhang, "Breast ultrasound image classification based on multiple-instance learning," *Journal of digital imaging*, vol. 25, no. 5, pp. 620–627, 2012.

[12] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[13] C.L. Zitnick and P. Dollár, "Edge boxes: Locating object proposals from edges," in *ECCV*. Springer, 2014, pp. 391–405.

[14] J.R. Uijlings, K.E. Van De Sande, T. Gevers, and A.W. Smeulders, "Selective search for object recognition," *IJCV*, vol. 104, no. 2, pp. 154–171, 2013.

[15] M.-M. Cheng, Z. Zhang, W.-Y. Lin, and P. Torr, "Bing: Binarized normed gradients for objectness estimation at 300fps," in *CVPR*, 2014, pp. 3286–3293.

[16] A. Krizhevsky, I. Sutskever, and G.E. Hinton, "Imagenet classification with deep convolutional neural networks," in *NIPS*, 2012, pp. 1097–1105.

[17] L. Pedraza, C. Vargas, F. Narvaez, O. Duran, E. Munoz, and E. Romero, "An open access thyroid ultrasound image database," in *ISMIPA*. International Society for Optics and Photonics, 2015, vol. 9287, p. 92870W.

[18] X. Xu and E. Frank, "Logistic regression and boosting for labeled bags of instances," in *PAKDD*. Springer, 2004, pp. 272–281.

[19] X.-S. Wei, J. Wu, and Z.-H. Zhou, "Scalable multi-instance learning," in *ICDM*. IEEE, 2014, pp. 1037–1042.

[20] X.-S. Wei and Z.-H. Zhou, "An empirical study on image bag generators for multi-instance learning," *Machine Learning*, vol. 105, no. 2, pp. 155–198, 2016.

[21] M. Hall, E. Frank, G. Holmes, B. Pfahringer, R. Reutemann, and I.H. Witten, "The weka data mining software: an update," *ACM SIGKDD explorations newsletter*, vol. 11, no. 1, pp. 10–18, 2009.